COMBINING SPATIALLY-EXPLICIT SIMULATION OF ANIMAL MOVEMENT AND EARTH OBSERVATION TO RECONCILE AGRICULTURE AND WILDLIFE CONSERVATION

Lei Song

MAY 2023

A DISSERTATION

Submitted to the faculty of Clark University,

Worcester, Massachusetts,

in partial fulfillment of

the requirements for

the degree of Doctor of Philosophy

in the Department of Geography

Dissertation Committee

Lyndon D. Estes, Ph.D.

Chief Instructor

John Rogan, Ph.D.

Committee Member

Christopher A. Williams, Ph.D.

Committee Member

Anna B. Estes, Ph.D.

Carleton College

Committee Member

Abstract

Agriculture will lead to significant biodiversity loss in Africa within the next few decades, as the region's agricultural systems adapt to meet rapidly growing food security challenges arising from rapid economic and population growth and climate change. Currently, agricultural expansion is the primary means of achieving growth in production. As expansion continues to rearrange our world, reconciling agricultural demands with the maintenance of biodiversity remains a critical challenge for conservation and livelihood. In this dissertation, I utilize a suite of Earth Observation (EO) products and cutting-edge modeling techniques to examine conservation strategies that try to balance between the demand for agriculture and the need for maintaining elephant habitat connectivity in a landscape fragmented by agriculture. In my first chapter, I introduce a novel mapping framework that enables the rapid generation of high-quality land cover maps by leveraging existing land cover products, followed by the training of a deep learning (DL) model for land cover mapping. I applied this approach to create a spatially detailed map of major land cover types in Tanzania, with a spatial resolution of 4.8 m. The high spatial resolution data obtained from this approach serves to fill critical gaps in ecosystem monitoring and conservation studies in Tanzania. In Chapter 2 and 3, I use the high-resolution land cover map created in Chapter 1, in combination with climatic, topographic, and anthropological data, to evaluate the environmental suitability and simulate habitat connectivity required for elephant survival. I identify critical corridors that are currently or will soon be occupied by agriculture and provide conservation recommendations to preserve essential habitat connectivity for

elephants, thereby supporting a viable population. In Chapter 4, I build upon the data and results from Chapters 1-3 and apply a land-use prioritization framework to optimize the allocation of agricultural land expansion. The objective is to identify regions suitable for expanding agriculture that can optimize crop production benefits while ensuring sustainable connectivity of elephant habitats, while also considering other ecological objectives. Through the use of this framework, it is demonstrated that with proper national management, we can maintain essential landscape connectivity while also meeting the food needs of the expanding population. However, agricultural productivity must improve to achieve this goal. The four chapters together demonstrate how progress in data, techniques, and theories provide both new opportunities and challenges in protecting species habitats and habitat connectivity in a rapidly changing world.

Academic History

Name: Lei Song **Date**: May 2023

Baccalaureate Degree:

Source:	**Date:**
B.S. Nanjing University of Information Science & Technology	June 2012

Master's Degree:

Source:	**Date:**
M.S. Nanjing University of Information Science & Technology	June 2015

Occupation and Academic Connection:	**Date:**
NASA FINESST Fellow	Sep 2020 – Aug 2023
Instructor, Clark University	Spring 2021
Research Assistant, Clark University	Jan 2018 – Dec 2019
Teaching Assistant, Clark University	Spring & Fall 2020, Fall 2017

Dedication

To the intertwined fate of all living beings, including wildlife and humanity.

Acknowledgements

This dissertation owes its completion to the guidance and support of so many, and any effort to fully acknowledge their help is likely to be inadequate. While words may prove insufficient, I would like to express my sincerest appreciation to all those who have played a role in making this dissertation possible.

First and foremost, I would like to thank the NASA Future Investigators in NASA Earth and Space Science and Technology (FINESST) program for providing financial support for my dissertation research. This opportunity has allowed me to pursue my academic goals and complete my doctoral degree. Also many thanks for the feedbacks and comments from anonymous award proposal reviewers.

I am deeply grateful to my advisor, Lyndon Estes, for his exceptional mentorship during my doctoral journey. He has not only taught me how to be a successful scientist through his guidance in teaching, writing, and research, but he has also been there for me during our shared struggles in the field. His dedication to science and unwavering support have provided me with a safe and supportive environment where I can grow and learn. His dedication and passion for science, evident in his tired but persistent face on many occasions, have inspired me to pursue my own research goals with enthusiasm and commitment. I feel incredibly fortunate to have had him as my advisor and to be his first PhD student, and I will always be grateful for his mentorship and guidance in helping me become the researcher I am today.

I would also like to thank my dissertation committee members Anna Estes, Christopher A. Williams, John Rogan for generously sharing their time and expertise to provide valuable feedback and insights that have contributed to the design of this dissertation and its subsequent funding. I would like to express my special gratitude to Anna Estes, whose profound field experience in Tanzania has enabled me to remotely sense that region. Lastly, thank them for their support in helping me successfully complete my doctoral exam and proposal defense during the first outbreak of the pandemic.

I am deeply thankful to all the individuals and organizations who provided the data, tools, and models used in this dissertation. Without their freely shared information and ideas, this dissertation would not have been possible.

I would like to express my deep appreciation to my family for their unconditional love, support, and encouragement. Thanks for their belief in me and always gentle check-ins on progress. I am truly grateful for all that they have done for me.

Finally, I express my heartfelt gratitude to my partner, Dr. Yifan Cai, for being an unwavering source of support and encouragement throughout my PhD journey. Your unwavering faith in me, your willingness to listen to my ideas, and your ability to offer insightful feedback have been invaluable to me. Your constant presence, both emotionally and physically, has helped me navigate the many challenges and setbacks that come with pursuing a PhD. I could not have made it this far without your help.

Table of Contents

List of Tables ..... **xv**

List of Figures ..... **xvi**

Introduction **1**

Tanzania: a country with critical large mammal diversity undergoing large-scale agricultural change 3

Land cover/land use mapping of smallholder-dominated savanna landscape 5

Elephant conservation within and between agricultural landscapes 7

Land use planning to balance agriculture and wildlife conservation 10

Promoting sustainable coexistence between wildlife and human communities 11

References 13

Chapter 1 ..... A super-ensemble approach to map land cover types with high resolution over data-sparse African savanna landscapes **21**

1.1 Abstract 21

1.2 Introduction 23

1.3 Materials and methods 27

1.3.1 Study area 27

1.3.2 Datasets 28

1.3.2.1 Satellite imagery 28

1.3.2.2 Pixel-level land cover (LC) reference labels 29

1.3.2.3 Land cover (LC) products 30

1.3.3 Modeling approach 31

1.3.3.1 Part 1: Ensemble multiple land cover (LC) products 31

1.3.3.2 Part 2: Creating gap-filled land cover labels 33
1.3.3.3 Part 3: Land cover classification with U-Net 35
1.3.4 Accuracy assessment 38
1.4 Results 39
1.4.1 Land cover label gap-filling 39
1.4.2 Land cover classification with U-Net 41
1.4.3 Final land cover map 45
1.5 Discussion 48
1.6 Conclusion 51
CRediT authorship contribution statement 52
Declaration of Competing Interest 52
Acknowledgments 52
References 53
Chapter 2 *itsdm*: Isolation Forest-based presence-only species distribution modeling and explanation in R 60
2.1 Abstract 60
2.2 Introduction 61
2.3 Package structure and description 63
2.4 SDMs with isolation forest 65
2.5 Application of Shapley values 67
2.5.1 Local explanation and applications in *itsdm* 67
2.5.2 Global explanation and applications in *itsdm* 69
2.6 Example 71
2.7 Comparison with other SDMs and recommendations 76

2.8 Discussion 79

Acknowledgements 80

Conflict of interests 81

Authors' contributions 81

Data accessibility statement 81

ORCID 81

References 82

Chapter 3 A national, multi-scale assessment of habitat connectivity of African savanna elephant (*Loxodonta africana*) 86

3.1 Abstract 86

3.2 Introduction 87

3.3 Materials and methods 90

3.3.1 Study area 90

3.3.2 African savanna elephant distribution data 92

3.3.3 Integrated multi-scale species distribution modeling 93

3.3.3.1 Regional species distribution modeling 96

3.3.3.2 Landscape species distribution modeling 97

3.3.3.3 Bayes Theorem-based map integration 99

3.3.4 Landscape connectivity 99

3.4 Results 101

3.4.1 Drivers of distributions and niches of African savanna elephants 101

3.4.2 Environmental suitability of African savanna elephants 104

3.4.3 Landscape connectivity and potential corridors 106

3.4.3.1 Connectivity between habitat clusters 106

3.4.3.2 Connectivity within habitat clusters 108

3.5 Discussion 111

3.5.1 Long-distance connectivity between primary habitat ecosystems 111
3.5.2 Small-range connectivity priorities and conservation recommendations 113
3.5.3 Conclusions and limitations 114
Data availability 115
Declaration of competing interest 115
Funding 116
References 117
Chapter 4 ..... Cropland allocation to minimize agriculture-elephant conflict with consideration of biodiversity and carbon costs 124
4.1 Abstract 124
4.2 Introduction 125
4.3 Materials and methods 128
4.3.1 Study region and objectives 128
4.3.2 Agricultural expansion and intensification 130
4.3.3 Elephant conservation 132
4.3.4 Biodiversity cost 133
4.3.5 Carbon cost 135
4.3.6 Trade-off model structure 137
4.4 Results 138
4.4.1 Status and effectiveness of protected areas (PAs) in Tanzania 138
4.4.2 Productive benefits and ecological costs for agricultural development in Tanzania 139
4.4.3 Contribution of landscape connectivity in land management 141
4.4.4 Significance of agricultural intensification 143
4.4.5 Harmonize different ecological criteria 145
4.5 Discussion 147

Acknowledgments 151

References 152

Conclusion 159

Advances in Earth Observation (EO) and Geospatial Artificial Intelligence (GeoAI) bring new opportunities as well as challenges for conservation studies 160

Elephants are confronted with unprecedented threats from human disturbances, but these can be mitigated with proper land management 162

Conservation requires forward-looking and multidimensional solutions 164

References 167

Appendix A Appendix to Chapter 1 170

A.1 PlanetScope NICFI basemap 170

A.2 Sentinel-1 Synthetic Aperture Radar (SAR) 171

A.3 Initial selection of Land cover (LC) products 172

A.4 U-Net structure and the relevant computation 175

A.5 Accuracy assessment metrics 176

A.6 Analysis of weight parameters of quality-weighted and class-balanced loss 178

A.7 Other supplementary tables and figures 180

References 183

Appendix B Appendix to Chapter 2 185

B.1 Supplementary tables, figures, and scripts 185

B.1.1 Supplementary tables and figures 185

B.1.2 Full version of the example 187

B.1.3 Scripts for other examples in the manuscript 203

References 203

B.2 Evaluation metrics 204
B.2.1 Presence-only evaluation 204
B.2.2 Presence-background evaluation 205
References 208

Appendix C Appendix to Chapter 3 210

C.1 Extended materials and methods 210
C.1.1 High-resolution land cover mapping in Tanzania 210
C.1.2 Multi-scale species distribution modeling 211
C.1.3 Landscape connectivity 219
C.2 Extended results and discussions 221
C.2.1 Species distribution modeling with polygon-based observations 221
C.2.2 Spatial scaling of environmental variables 223
C.2.3 Supplementary figures and tables 229
References 232

Appendix D Appendix to Chapter 4 234

D.1 Agricultural yield data 234
D.2 Elephant conservation data 237
D.3 Biodiversity data 238
D.4 Carbon data 241
References 242

List of Tables

Table 1-1. Land cover products used and their accuracies over the study region. 30
Table 1-2. The evaluation of the random forest gap-filling model..... 39
Table 1-3. Validation, independent test accuracy and prediction confidence of U-Net model 42
Table 2-1 Core functions and descriptions in *itsdm*..... 64
Table 3-1 Selected environmental variables for the regional and landscape species distribution model. 96
Table 4-1. Agricultural and ecological statistics inside and outside of protected areas (PAs) in Tanzania. Plant area is the overall area in hectares. Other values, proactive biodiversity index (BIp), carbon density, and elephant migration index (EI), use the mean values. 139
Table 4-2. New allocated cropland and ecological costs under different scenarios. 143
Table A-1. The conversion table from original classes to classes defined in this study..... 174
Table A-2. Evaluation of different LC products using the independent test dataset 175
Table A-3. Estimated time spent by one interpreter on label creation, comparing time required for a hypothetical case in which tiles were labelled manually based on visual image interpretation to the estimated time spent editing labels using the semi-automated labelling approach developed in this study..... 180
Table B-1. Decisive arguments setting to specify the model type..... 185
Table C-1. Variables hypothesized to affect African savanna elephant distribution in Tanzania. 230
Table D-1. Evaluation of current and attainable crop yield downscaling models 235
Table D-2. Information of species included in the analysis* 240

List of Figures

Figure 1-1. The ecozones of the study area overlaid by the PlanetScope basemap tiling grid 27
Figure 1-2. Seasonal NICFI basemap in false-color, harmonic coefficients (RGB: Slope, $\cos\left(\frac{2\pi t}{d_{yr}}\right)$, Intercept) of Sentinel-1 dB in VV and VH polarization, spectral signature, and temporal signature of land cover samples in an example tile (1227-1002) 29
Figure 1-3. The workflow of the proposed approach (OSM stands for OpenStreetMap) 31
Figure 1-4. Sub-tiling system and different types of label examples (A are the weak ensemble labels, B are the gap-filled labels produced by a Random Forests model, C are the human refined labels, and S1 and S2 are false-color composites of PlanetScope NICFI basemaps in season 1 and season 2) 35
Figure 1-5. Distribution of human refined label quality 41
Figure 1-6. Evolution of the learning rate (A), loss function (B), mean IOU (C), accuracy of each class (D1-D7), and average accuracy (D8)) on the validation dataset during model training 43
Figure 1-7. Confusion matrix heatmap of the independent test of U-Net classification model 44
Figure 1-8. The confidence map of land cover classification by U-Net 45
Figure 1-9. Predicted land cover (A), error-adjusted area estimates (B), and proportion of each land cover type (C) with 95% confidence interval over the study area (*The estimated area of wetland is directly calculated by the rasterized OSM layer without error adjustment) 46
Figure 1-10. Seasonal NICFI images, prediction in this study, and the CGLS_LC100m and FROM-GLC 2017v1 product of four example tiles (a and b are in Central plateau agro-ecological zone, c is in Northern Riftzone and Volcanic highlands zone, and d is in Eastern plateau zone) 47

Figure 2-1. Schematic representation of a single tree (a) and its feature space (b) for an Extended Isolation Forest (EIF) built by a two-dimensional dataset..... 66
Figure 2-2. Environmental change analysis of BIO1 (annual mean temperature) to Za Baobab tree (*Adansonia za Baill.*) in Madagascar. Panel A shows that Za Baobab tree has a positive linear response to annual mean temperature in Madagascar. 23.28℃ is the tipping point of annual mean temperature, which means the Za Baobab tree in areas with an annual mean temperature near 23.28℃ is vulnerable to a cooling temperature. Panel B shows Za Baobab tree in most areas of Madagascar will not be affected by a changing annual mean temperature. The annual mean temperature in Northwest coastal areas will become not suitable for Za Baobab tree. 71
Figure 2-3. Variable importance of virtual species case diagnosed by Shapley values technique. Variables bio12, bio5, and bio1 have much higher importance than var 1 through var 3, as intended. In addition, the similarity in the values for these metrics for both the training and test dataset indicates that the model is generalizable. 73
Figure 2-4. Shapley value-based response curves of bio1 and bio12 colored by bio5 in our virtual species case. The modeled species has a strong positive response to both bio1 and bio12 that respectively peak at 25 °C and 1000 mm, and that the two are also strongly correlated with bio5, particularly in the upper range for bio1 and in the lower to mid-range for bio12. 74
Figure 2-5. Shapley values-based spatial response map of variable bio12 in our virtual species case. It is evident that bio12 contributes minimally in some areas even though it is the most vital environmental variable diagnosed in variable analysis. 75
Figure 2-6. Variable contributions to the modeled suitability of an occurrence observation. 76
Figure 2-7. Performance comparison between Isolation Forest (iForest) and other mainstream SDM models. Evaluation metrics are Area Under the ROC Curve (AUC) , Pearson correlation between modeled and real habitat suitability (COR), True Skill Statistic with a threshold of 0.5 ($TSS_{0.5}$), and Euclidean distance between modeled and real suitability. Normal detection type is the case that the species can be detected in any areas with suitability higher than 0.5 or 0.6 and Core area type is the case that the species can only be detected in

areas with suitability higher than 0.8 or 0.9. The figure in upright is an example drawn with No.6 virtual species. 78
Figure 3-1. The study area and population observations (Expert range map, census blocks, and occurrences obtained from GBIF database) of African savanna elephants. Spatially-thin region delineates the region with overdense biased occurrences that need to be spatially thinned. (See details in section 3.3.3.2). 91
Figure 3-2. The integrated approach for multi-scale species distribution modeling. SDM means species distribution modeling. The var. is the abbreviation of variables. 95
Figure 3-3. Variable importance at different coarse scales (10, 5, and 2.5 arc-minutes) for regional species distribution model (SDM). The mean of absolute Shapley values (x-axis) is used to indicate the importance of a variable in the SDM, with higher values indicating greater significance. 103
Figure 3-4. Variable importance at fine scale (1 km) for landscape species distribution model (SDM). #km is the window size of focal statistics. The mean of absolute Shapley values (x-axis) is used to indicate the importance of a variable in the SDM, with higher values indicating greater significance. 104
Figure 3-5. Environmental suitability map at the regional scale (5 arc-minutes, a), the landscape scale (0.5 arc-minutes, b), and Bayes fusion (c) 105
Figure 3-6. Landscape connectivity between primary habitat clusters (1, 2, & 3) under scenarios A (a) and B (b). In scenario A, pixels with >80% cropland coverage and containing >400 settlements are considered as barriers for elephant movement. In scenario B, pixels with >20% cropland coverage and containing >200 settlements are considered as potential barriers for elephant movement assuming current human disturbances will keep expanding over these areas. Areas Ⅰ, Ⅱ, Ⅲ, and Ⅳ are highlighted areas. 108
Figure 3-7. Difference in predicted landscape connectivity within each primary habitat cluster (No. 1-3 in Figure 3-6) under scenario A and B (section 3.3.4) and the detected critical regions of African elephant conservation in Tanzania. The current density changes were calculated by dividing the current density predicted under scenario B by the current density predicted under scenario A (Figure C-9 in Appendix C). The bluish color indicates

areas with decreasing usage by elephants with the expansion of human activities, and the coral color indicates areas with increasing usage by elephants for movement, and thus become more critical for landscape connectivity. 110
Figure 4-1. Spatial location and geographic characteristics of Tanzania. 129
Figure 4-2. Decision-making factors of each planning unit in land allocation analysis: attainable production (a), travel time (b), biodiversity index (c), carbon density (d), and elephant migration index (e). Attainable food crop production (and gain) and carbon cost are obtained with the scenario of 60% of each planning unit (100 ha) being allowed to be cultivated 140
Figure 4-3. New selected agricultural areas and ecological costs under different solutions: (a) only considering agricultural benefits (production gain and transport time), (b) considering both agricultural benefits and ecological costs (biodiversity cost and carbon cost), and (c) considering agricultural benefits, ecological costs (biodiversity cost, carbon cost, and elephant migration cost). Attainable food crop production is evaluated assuming that 60% of each planning unit (100 ha) can be cultivated (section 4.3.6). 142
Figure 4-4. Comparison between current cropland and newly allocated cropland under 80% (a), 60% (b), 80% with half yield gain (c), and 60% with half yield gain (d) scenario. Current cropland includes all planning units with cultivated area. #% indicates the allowed maximum percentage of each planning unit (100 ha) to expand for cultivation. Planning units currently with cultivated areas higher than #% would not be reduced. Equal weights were set across all objectives. Note that with half yield gap closed, current spare land can no longer guarantee food demand 145
Figure 4-5. Tradeoff surface between biodiversity cost, carbon cost and elephant migration cost using different weights in the decision-making (a), and agricultural land expansion with the highest balance between ecological costs (b). The tradeoff surface was generated using a fixed weight (25%) for the agricultural benefit, which was equally split between current production, attainable production gain, and transport time. So the weight for each analyzed ecological factor ranges from 0 to 75%. The values of three ecological costs were normalized

between their minimum and maximum values. The balance score is the standard deviation of three cost values. 147
Figure A-1. Selected Land Cover products we evaluated in pre-assessment (Note: Esri Land Cover only has rangeland, which is colored the same as shrubland in the other two maps. In GFSAD +TanSIS map, orange is their overlap area, red is the area covered by TanSIS only, and yellow is the area covered by GFSAD30 only) 174
Figure A-2. Illustration of the U-Net structure for semantic segmentation (modified from (Ronneberger et al., 2015)) 176
Figure A-3. Sensitivity analysis on weight parameters of quality-weighted and class-balanced loss 179
Figure A-4. Variable importance and tuning curve of random forest guessing model (S1 stands for NICFI season 1 image, S2 stands for NICFI season 2 images. B, G, R, and NIR are four spectral bands and NDVI, EVI, SAVI, and ARVI are four vegetarian indices. VV and VH are two types of polarization of Sentinel-1 images. Intercept, slope, and coefs are coefficients of harmonic regression fitting. sin/cos(t) represents one intra-annual seasonal cycle and sin/cos(2t) describes two intra-annual cycles. See more details in Section 1.3.2.1, A.1 and A.2)..... 181
Figure A-5. Confusion matrix heatmap of the independent test of Random Forests model 182
Figure B-1.Environmental suitability maps of No.6 virtual species predicted by different SDMs using different types of samples 186
Figure B-2. Environmental suitability and response curves of the simulated virtual species. 189
Figure B-3. Outlines detected in the occurrence dataset..... 192
Figure B-4. The model diagnostic shown by plot function (made by `plot(mod$eval_test)`) 196
Figure B-5. Variable importance analysis of species distribution model for the virtual species. 197
Figure B-6. Marginal response curves of bio1 and bio12 for the virtual species. 199
Figure B-7. Spatial response map of variable bio12 for the virtual species. 200

Figure B-8. Variable contributions to the modelled suitability of selected occurrence observations. 201
Figure B-9. Presence-absence conversion maps. 202
Figure C-1. Kolmogorov–Smirnov (K-S) distances between values of environmental variables extracted by occurrences and expert range map. VRM is Vector Ruggedness Measurement; ED is edge density; AREA_MN is mean of patch area; CONTIG_MN is mean of contiguity index; PD is patch density. 217
Figure C-2. Examples of pixels (1 km^2) with over 80% cropland coverage (a, 83%), containing more than 400 settlements (b, 465/km^2), with over 20% cropland coverage (c, 22%), and containing more than 200 settlements (d, 237/km^2). Panel a and b represent the landscape that already have intensive human disturbances, and panel c and d represent the landscape that potentially will become intensively disturbed by human activities. 219
Figure C-3. Three primary habitat clusters used for landscape connectivity analysis and the involved Protected areas (PAs) in Tanzania. 220
Figure C-4. Effect of sampling ratio of pseudo-occurrences on model performance evaluated by different evaluation metrics (mean and 95% confidence intervals) based on cross validation at different coarse scales (2.5, 5, and 10 arc-minutes). 222
Figure C-5. Comparison of model performance evaluated by different evaluation metrics (mean and 95% confidence intervals) based on cross validation at different coarse scales (2.5, 5, and 10 arc-minutes). A is the evaluation based on cross validation using pseudo-occurrences with the best sampling ratio (section 3.3.3.1 and Figure 3-2). B is the sensitivity calculated by real occurrences with the optimal threshold obtained from cross validation. 223
Figure C-6. Statistical responses of African savanna elephants to environmental variables of type A. There is no evident scale dependence for this type. Variables along the x-axis with a superscript ^{L} were $\log_{10}$ transformed. The differences in residual values of seasonal NDVI, NDVI seasonality, and BIOs mainly result from the different grain sizes of doing aggregated calculations (section 3.3.3.1 in the main text). 226

Figure C-7. Statistical responses of African savanna elephants to environmental variables of type B. Response curves have the same central values but different ranges across scales. Variables along the x-axis with a superscript ^{L} were $\log_{10}$ transformed 227
Figure C-8. Statistical responses of African savanna elephants to environmental variables of type C. Response curves have similar shapes but horizontal shifts across different scales. Variables along the x-axis with a superscript ^{L} were $\log_{10}$ transformed 228
Figure C-9. Landscape connectivity within each primary habitat cluster (No. 1-3 in Figure 3-6 in the main text) under scenario A and B (section 3.3.4) and the detected critical regions of African elephant conservation in Tanzania. 229
Figure D-1. Current yield map of maize, rice, cassava, and pulses at 5 minutes (A) and the downscaled results at 1 km (B) 236
Figure D-2. Attainable yield map of maize, rice, cassava, and pulses at 5 minutes and the downscaled results at 1 km 237
Figure D-3. Environmental suitability of elephants and census blocks (shown in white) used in Circuitscape 238
Figure D-4. Weighted species richness and rarity-weighted species richness 240
Figure D-5. MSA, BII, and BHI used in the analysis 241
Figure D-6. Aboveground biomass, belowground biomass, and soil carbon stocks in the top 1 m 241

Introduction

Africa south of the Sahara (SSA) is poised to become the 21st Century's hotspot of land use change, as agriculture expands to meet the food demands of the region's rapidly growing populations and economies (Chaplin-Kramer et al., 2015; Laurance et al., 2014; Lindenmayer et al., 2012; Searchinger et al., 2015). SSA is also a critical region for biodiversity, as it hosts some of the world's last remaining megafaunal populations (e.g. African elephants, Loxodonta Africana) whose behaviors support a broad range of other species (Tilman et al., 2017; Vries et al., 2018). Agriculture is the leading driver of terrestrial biodiversity loss throughout the world (Kehoe et al., 2017; Tilman et al., 2017), and particularly threatens these large mammals and their ecosystems in SSA (Laurance et al., 2014; Searchinger et al., 2015). Without appropriate management of agricultural land-use, agricultural expansion might lead to disastrous consequences for some of the world's most important and unique biological diversity (Tilman et al., 2017; Vries et al., 2018).

While some consider biodiversity loss to be an irreversible consequence of human disturbances, effective land management practices can help reduce the impact of agricultural expansion and potentially restore some of the losses (Tilman et al., 2017; Vries et al., 2018). In recent years, substantial attention has been placed on the nexus between agriculture and conservation, with a particular focus on finding balances between agricultural and environmental needs (Estes et al., 2016; Glamann et al., 2017; Shackelford et al., 2015). Two fundamentally different approaches have received attention widely: land sparing that

integrates conservation and food production on the same land; and land sharing that separates land for conservation from land for crops (Jóhannesdóttir et al., 2017). Yet in this era, as the nature of agriculture is rapidly changing, the tools and techniques used to develop these management practices need to be dynamically evaluated and updated to keep pace. For instance, neither land sharing nor land sparing guarantees co-benefits of conservation and agriculture without sufficient geospatial information on habitat and habitat connectivity, especially large mammals (Phalan et al., 2011). Large animals prefer certain terrain, traverse select corridors at fixed times of year, and seek particular ecological conditions (Kays et al., 2015). Despite occurring outside of carefully designed protected areas (PAs), in SSA, intensive agriculture has been a significant barrier to the interconnection of fragmented habitats, resulting in reduced effective population size, limited gene flow between populations, and increased human-wildlife conflicts (Lohay et al., 2020).

Knowledge of landscape connectivity and its associated challenges will help to identify lands that are essential for wildlife survival, more suited for expanding agriculture, and shareable for both but in ways that help to reduce crop-wildlife conflict (Crespin & Simonetti, 2018; Hill et al., 2015; Phalan et al., 2011). Remote sensing (RS) can play a vital role in quantifying ecosystem processes and states at various spatio-temporal scales, particularly in the data-sparse regions of Africa. Additionally, RS can help develop tools that inform land-use planning to achieve a balance between biodiversity conservation and human livelihood objectives.

In my dissertation, I use a suite of satellite imagery and Earth Observation (EO) products to investigate conservation strategies that achieve a balance between the demand for

agriculture and the need for wildlife habitat connectivity, while also considering other ecological priorities in a landscape fragmented by agriculture. In chapter 1, I combine high-resolution RS products and Geospatial Artificial Intelligence (GeoAI) techniques to generate a detailed map of a complex savanna landscape, differentiating between areas of agriculture and natural habitats. Relying on this map, in Chapter 2 and 3, focusing on elephants, a keystone species, I examine the impacts of human activities on their habitat and habitat connectivity, and provide conservation recommendations. Finally, in Chapter 4, I focus on identifying the optimal areas for future agricultural expansion that have minimal impacts on elephant habitat connectivity, while also considering other conservation goals. Through my research, I found that it is feasible to achieve a balance between agriculture and elephant habitat connectivity, as one of conservation objectives, in a mosaic landscape, and that RS technologies can play a critical role in informing conservation strategies.

Tanzania: a country with critical large mammal diversity undergoing large-scale agricultural change

The United Republic of Tanzania is the largest country located in Eastern Africa, covering an area of approximately 947,300 km^2 (Tanzania National Bureau of Statistics, 2021), and has a diverse geography and climate (Luhunga et al., 2018). It is a country within the African Great Lakes region in East Africa. A plateau averaging 900 - 1,800 m in height makes up the greater part of the country (Tanzania National Bureau of Statistics, 2021). The western and eastern branches of the East African Rift System (EARS) form a complex topography in Tanzania, generating substantial variability in climate over space (Rowhani et

al., 2011) that corresponds to a broad range of primary ecosystems including savanna, tropical and subtropical forest, and montane (Burgess et al., 2004; John et al., 2020), which collectively support globally significant biodiversity (Nkwabi et al., 2018). It also has one of the largest remaining large mammal migrations, which make it a critical area for maintaining habitat connectivity (Metzger et al., 2015). Approximately a third of the country's territory is protected across 22 national parks and other management areas, providing core habitat for scores of animals, including elephants (Tanzania National Bureau of Statistics, 2021). It is therefore known for various species and high biodiversity, but it also faces increasing challenges in balancing population increase and wildlife conservation (Giliba et al., 2022).

Agriculture is the primary source of food and the largest contributor to the national economy, but its growth is slow (van Ittersum et al., 2016). In Tanzania, as in other sub-Saharan countries, smallholder farmers (0.2 to 3.0 hectares) dominate agriculture, relying mainly on hand hoes and rain-fed agriculture, and occupying over 80 percent of the arable land (FAO, 2019). Improving agricultural-based livelihoods and food security is a key objective of the national strategies for growth and the reduction of poverty (Suleiman, 2018). Despite efforts, agriculture has not been able to act as an engine for growth, economic transformation, and poverty reduction in Tanzania. This is primarily due to production constraints faced by smallholder farmers, such as the small sizes of their farms, insufficient access to modern agricultural techniques, low income levels, inadequate policies, and insufficient government support in the form of subsidies for agricultural inputs (Suleiman, 2018).

Elephants, as one of the largest and most widely distributed species in Tanzania, are largely threatened by habitat loss, fragmentation, and degradation, and often in conflict with local communities over crop damage and destruction of property (Shaffer et al., 2019). The challenges of balancing agricultural development and elephant conservation are complex and require a holistic approach that involves the engagement of local communities, government agencies, and other stakeholders (Dickman, 2010; Hoare, 2015; Shaffer et al., 2019). A balance is in need to ensure both sustainable agricultural development and the conservation of elephants and other wildlife species in Tanzania.

Land cover/land use mapping of smallholder-dominated savanna landscape

The creation of a land cover map is typically the initial and fundamental step in landscape analysis (Olofsson et al., 2012). The spatial extent and resolution of the map determine the potential depth of analysis, and the accuracy of the map influences the reliability of subsequent assessments (Sawaya et al., 2003; Strahler et al., 2006). Therefore, the production of large-scale, timely, and precise land cover maps is essential for various environmental applications, including biodiversity conservation, natural resource management, and food security assessment (Anderson et al., 2017; Jin et al., 2019; Leite-Filho et al., 2021; Pettorelli et al., 2016; X.-P. Song et al., 2018). In conservation analysis, for example, a high spatial resolution is advantageous in assessing the local impacts of human activities on ecosystems and accounting for the close associations that many species have with environmental features (e.g. human disturbances) that can change over short distances

(Mokany et al., 2020). However, current land cover maps are often inadequate for agricultural and ecological applications, particularly over data-sparse regions, such as SSA, due to their limited spatial and temporal coverage and low accuracy (Gómez et al., 2016).

Recent advances in EO, such as high spatial/temporal resolution satellites PlanetScope and Sentinel-1 (Planet Team, 2017; Roy et al., 2021; Torres et al., 2012), and improvements in GeoAI techniques (Janowicz et al., 2020), present novel opportunities for land cover mapping in SSA. However, the need to collect extensive and task-specific training and reference datasets poses a significant hurdle to a wider application of these datasets and techniques. Although recent investments have been made to develop global ground truth datasets (Burke & Lobell, 2017; Laso Bayas et al., 2017; Schmitt et al., 2019), the available data typically prove inadequate for training qualified GeoAI models over large areas. The laborious and time-intensive process of collecting labelled training data necessitates a more efficient labeling strategy to meet the demand for high-resolution land use/land cover products in those data-sparse but highly dynamic landscapes.

In my first chapter, I develop an efficient, semi-automated approach that trains a deep learning (DL) model capable of generating improved, high resolution land cover data in hard-to-map, data sparse landscape. The approach involves the generation of synthetic labels by extracting overlapped parts from several existing land cover products (Buchhorn, Lesiv, et al., 2020; Buchhorn, Smets, et al., 2020; Congalton et al., 2017; Xu et al., 2019), which provides an initial set of labels but with partial coverage of the study region (excluding areas lacking consensus). Subsequently, I train a Random Forest model using predictors based on temporal features extracted from Sentinel-1 time series, along with raw bands of the

PlanetScope basemaps and derived vegetation indices, to fill in the missing data in the synthetic labels for selected tiles. The resulting gap-filled labels undergo assessment and editing and are then used to train a U-Net land cover model.

As demonstrated by a pilot experiment conducted in Northern Tanzania, this approach proves to be an effective method for mapping complex landscapes. Afterwards, I applied this approach to the entire country, producing a high-quality land cover map of Tanzania that includes 8 major land cover types and has a spatial resolution of 4.77 m. This map has the potential for numerous applications and serves as the foundation for the subsequent analysis conducted throughout this dissertation. For instance, in Chapter 3, the high-resolution land cover map allows for the consideration of complex landscape structure in small windows (e.g. 1 km) when conducting species distribution modeling. In Chapter 4, the estimation of the cropland area from the map, which has higher boundary precision and a finer resolution, provides a more accurate assessment of the current cultivated areas and those likely to experience future agricultural expansion.

Elephant conservation within and between agricultural landscapes

Following the acquisition of a dependable, high-resolution land cover map, my dissertation delves deeper into the spatial understanding of the habitat and habitat connectivity of African savanna elephants in Tanzania, a landscape shaped by agriculture. African savanna elephants, being a keystone species, are instrumental in shaping natural ecosystems and facilitating a diverse range of other species (Kohi et al., 2011). Modeling

their habitat and habitat connectivity can therefore have critical ecological meaning, and also validate tools and techniques for species distribution modeling (SDM). It is therefore not surprising that significant efforts have been invested in scrutinizing the habitats and habitat connectivity of African elephants (Bastille-Rousseau et al., 2020; Chamaillé-Jammes et al., 2007; Cisneros-Araujo et al., 2021; Jones et al., 2012; Lohay et al., 2020; Shaffer et al., 2019; Zacarias & Loyola, 2018). Nonetheless, owing to the species' large range and multiple restrictions, such as the challenges of data collection and limited resources, spatially explicit studies are often restricted to small areas, such as national parks, and require extensive data to develop and validate (Bukombe et al., 2022; de Knegt et al., 2011; Douglas-Hamilton et al., 2005; Martin et al., 2019; Schüßler et al., 2018; Tshipa et al., 2017). Studies conducted at a broader scale typically tend to overlook or neglect the finer details of the landscape (Chase et al., 2016; Dejene et al., 2021; Robson et al., 2017; Wall et al., 2021). Past research (Foley & Foley, 2022; Jones et al., 2012; Lohay et al., 2020; Ntukey et al., 2022) demonstrates that human activities have already restricted or impeded corridors between some of their habitats that were previously extensively utilized. Therefore, to safeguard one of the world's most important species and the associated unique biological diversity, it is imperative to assess their spatially explicit habitat quality and habitat connectivity status and associated risks beyond the borders of formally protected areas at a national level or higher, despite the difficulties involved.

In Chapter 2 and 3, I employ a multi-scale integrated approach to evaluate the geospatial suitability of habitat for African savanna elephants in Tanzania, with a particular emphasis on the spatial scaling of environmental predictors. Using a hierarchical approach to

consider environmental features associated with elephants at different levels enables me to consider the constraints at the higher levels and the mechanisms at the lower levels that determine elephant distribution in the landscape (de Knegt et al., 2011; Turner & Gardner, 2015). The multi-scale structure also allows me to integrate multiple data sources, which differ in terms of design and accuracy, such as the expert range map and the occurrence survey dataset. To mitigate sampling bias issues in species distribution modeling (SDM), in chapter 2, I implement an Isolation Forest-based model (Liu et al., 2008, 2012; L. Song & Estes, 2023), which is a widely used anomaly detection algorithm in computer science, and develop an R package. Next, I use the resulting spatial suitability as input to Circuitscape, a graph-based landscape connectivity model (Brad H. McRae et al., 2008), to calculate the importance of a land unit in maintaining necessary habitat connectivity. Setting different densities of agriculture and settlement as movement barriers provides me with insights into how human activities impact the communication between populations in separate habitats, and helps to identify which corridors have already been blocked by human activities and may need restoration, and which corridors are at the risk of disappearing if surrounding human activities continue to expand without proper management. Based on the simulated results, my coauthors and I propose a set of policy recommendations for elephant conservation in Tanzania to maintain healthy flows between habitats.

Land use planning to balance agriculture and wildlife conservation

After identifying the areas where landscape connectivity is most severely affected by human activity, my dissertation continues to find solutions to mitigate these impacts and optimize the coexistence of elephants and humans in a future scenario. As the competition for space on our planet continues to increase, land-use prioritization has become an increasingly significant tool to make difficult decisions about which places and species to protect with limited resources. Particularly in the context of agricultural development, these tools can be used to identify areas for intensifying crop production, expanding new agricultural lands while minimizing the impact on biodiversity. Chapter 3 has shown that elephants are still at risk from human activities even with well-established PAs because their movements can extend beyond the boundaries of formally protected areas. Therefore, in Chapter 4, I expand on an existing land-use prioritization framework (Estes et al., 2016) to reconcile the demands of humans and biodiversity by including large-scale and spatially explicit landscape connectivity, with the goal of increasing the effectiveness of conservation efforts.

Land use planning is a complex decision-making process that requires considering multiple criteria and directions to integrate various objectives among stakeholders, even though sometimes specific objectives may be prioritized (Estes et al., 2016; Phalan et al., 2011). Thus, in my land-use prioritization model, I incorporate other ecological and agricultural factors along with landscape connectivity to facilitate forward-looking land

allocation, maximizing productive benefits while minimizing ecological costs. To quantify these factors, I use spatially explicit metrics and indices. Specifically, I calculate the attainable production of a land unit and the difference between attainable and current production to evaluate its agricultural potential in the future. Additionally, I integrate multiple biodiversity metrics at both species and ecosystem levels to generate a proactive biodiversity index that highlights the biodiversity value of a given land unit. Finally, I estimate the vegetation biomass and soil organic carbon to determine the possible carbon release from land conversion.

Placing varying weights on different factors in a linear transformation, a solution is generated that highlights different ecological and agricultural objectives, including identifying which current croplands can be kept and intensified and which areas have best potential for agriculture expansion. My analysis offers valuable insights into the quantitative trade-offs between the different factors by assigning varying weights to each of them.

Promoting sustainable coexistence between wildlife and human communities

With increasing conflicts between expanding agriculture and diversity, this dissertation aims to investigate a sustainable solution that improves coexistence between wildlife and human communities. All four chapters, which utilize geospatial analysis and modeling, together demonstrate that the establishment of large protected areas based solely on prior conservation assessments is not the end of story. Instead, biodiversity conservation must be pursued through geospatially context-dependent designation that considers both

direct demands, such as habitat requirement, and indirect needs, such as continuous gene flow, along with human livelihoods.

References

Anderson, S. J., Ankor, B. L., & Sutton, P. C. (2017). Ecosystem service valuations of South Africa using a variety of land cover data sources and resolutions. *Ecosystem Services*, *27*, 173–178. https://doi.org/10.1016/j.ecoser.2017.06.001

Bastille-Rousseau, G., Wall, J., Douglas-Hamilton, I., Lesowapir, B., Loloju, B., Mwangi, N., & Wittemyer, G. (2020). Landscape-scale habitat response of African elephants shows strong selection for foraging opportunities in a human dominated ecosystem. *Ecography*, *43*(1), 149–160. https://doi.org/10.1111/ecog.04240

Brad H. McRae, Brett G. Dickson, Timothy H. Keitt, & Viral B. Shah. (2008). Using Circuit Theory to Model Connectivity in Ecology, Evolution, and Conservation. *Ecology*, *89*(10), 2712–2724.

Buchhorn, M., Lesiv, M., Tsendbazar, N.-E., Herold, M., Bertels, L., & Smets, B. (2020). Copernicus global land cover layers—Collection 2. *Remote Sensing*, *12*(6), 1044.

Buchhorn, M., Smets, B., Bertels, L., De Roo, B., Lesiv, M., Tsendbazar, N.-E., Herold, M., & Fritz, S. (2020). Copernicus Global Land Service: Land Cover 100m: Collection 3 Epoch 2018, Globe. *Version V3. 0.1)[Data Set]*.

Bukombe, J., Marealle, W., Kimaro, J., Kija, H., Kavana, P., Kakengi, V., Nindi, J., Keyyu, J., Ntalwila, J., Kilimba, N., Bwenge, F., Nkwabi, A., Lowassa, A., Sanare, J., Mwita, M., Leweri, C., Kohi, E., Mangewa, L., Juma, R., … Lobora, A. (2022). Viability assessment of the Wami-Mbiki Game Reserve to Nyerere National Park wildlife corridor in southern Tanzania. *Global Ecology and Conservation*, *39*, e02259. https://doi.org/10.1016/j.gecco.2022.e02259

Burgess, N., Hales, J., Underwood, E., Dinerstein, E., Olson, D., Itoua, I., Schipper, J., Ricketts, T., Newman, K., & others. (2004). *Terrestrial ecoregions of Africa and Madagascar: A conservation assessment.* Island Press.

Burke, M., & Lobell, D. B. (2017). Satellite-based assessment of yield variation and its determinants in smallholder African systems. *Proceedings of the National Academy of Sciences*, *114*(9), 2189–2194. https://doi.org/10.1073/pnas.1616919114

Chamaillé-Jammes, S., Valeix, M., & Fritz, H. (2007). Managing heterogeneity in elephant distribution: Interactions between elephant population density and surface-water availability: Surface water and elephant distribution. *Journal of Applied Ecology*, *44*(3), 625–633. https://doi.org/10.1111/j.1365-2664.2007.01300.x

Chaplin-Kramer, R., Sharp, R. P., Mandle, L., Sim, S., Johnson, J., Butnar, I., Milà i Canals, L., Eichelberger, B. A., Ramler, I., Mueller, C., McLachlan, N., Yousefi, A., King, H., & Kareiva, P. M. (2015). Spatial patterns of agricultural expansion determine

impacts on biodiversity and carbon storage. *Proceedings of the National Academy of Sciences*, *112*(24), 7402–7407. https://doi.org/10.1073/pnas.1406485112

Chase, M. J., Schlossberg, S., Griffin, C. R., Bouché, P. J. C., Djene, S. W., Elkan, P. W., Ferreira, S., Grossman, F., Kohi, E. M., Landen, K., Omondi, P., Peltier, A., Selier, S. A. J., & Sutcliffe, R. (2016). Continent-wide survey reveals massive decline in African savannah elephants. *PeerJ*, *4*, e2354. https://doi.org/10.7717/peerj.2354

Cisneros-Araujo, P., Ramirez-Lopez, M., Juffe-Bignoli, D., Fensholt, R., Muro, J., Mateo-Sánchez, M. C., & Burgess, N. D. (2021). Remote sensing of wildlife connectivity networks and priority locations for conservation in the Southern Agricultural Growth Corridor (SAGCOT) in Tanzania. *Remote Sensing in Ecology and Conservation*, *7*(3), 430–444. https://doi.org/10.1002/rse2.199

Congalton, R., Yadav, K., McDonnell, K., Poehnelt, J., Stevens, B., Gumma, M., Teluguntla, P., & Thenkabail, P. (2017). *Global Food Security-support Analysis Data (GFSAD) Cropland Extent 2015 Validation 30 m V001*.

Crespin, S. J., & Simonetti, J. A. (2018). Reconciling farming and wild nature: Integrating human–wildlife coexistence into the land-sharing and land-sparing framework. *Ambio*, 1–8. https://doi.org/10.1007/s13280-018-1059-2

de Knegt, H. J., van Langevelde, F., Skidmore, A. K., Delsink, A., Slotow, R., Henley, S., Bucini, G., de Boer, W. F., Coughenour, M. B., Grant, C. C., Heitkönig, I. M. A., Henley, M., Knox, N. M., Kohi, E. M., Mwakiwa, E., Page, B. R., Peel, M., Pretorius, Y., van Wieren, S. E., & Prins, H. H. T. (2011). The spatial scaling of habitat selection by African elephants: Scaling habitat selection by elephants. *Journal of Animal Ecology*, *80*(1), 270–281. https://doi.org/10.1111/j.1365-2656.2010.01764.x

Dejene, S. W., Mpakairi, K. S., Kanagaraj, R., Wato, Y. A., & Mengistu, S. (2021). Modelling continental range shift of the African elephant (*Loxodonta africana*) under a changing climate and land cover: Implications for future conservation of the species. *African Zoology*, *56*(1), 25–34. https://doi.org/10.1080/15627020.2020.1846617

Dickman, A. J. (2010). Complexities of conflict: The importance of considering social factors for effectively resolving human-wildlife conflict: Social factors affecting human-wildlife conflict resolution. *Animal Conservation*, *13*(5), 458–466. https://doi.org/10.1111/j.1469-1795.2010.00368.x

Douglas-Hamilton, I., Krink, T., & Vollrath, F. (2005). Movements and corridors of African elephants in relation to protected areas. *Naturwissenschaften*, *92*(4), 158–163. https://doi.org/10.1007/s00114-004-0606-9

Estes, L. D., Searchinger, T., Spiegel, M., Tian, D., Sichinga, S., Mwale, M., Kehoe, L., Kuemmerle, T., Berven, A., Chaney, N., Sheffield, J., Wood, E. F., Caylor, K. K., & others. (2016). Reconciling agriculture, carbon and biodiversity in a savannah

transformation frontier. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1703), 20150316. https://doi.org/10.1098/rstb.2015.0316

FAO. (2019). *The United Republic of Tanzania Resilience Strategy 2019–2022*. Food and Agriculture Organization of the United Nations Rome, Italy.

Foley, C. A., & Foley, L. S. (2022). The History, Status, and Conservation of the Elephant Population in the Tarangire Ecosystem. In *Tarangire: Human-Wildlife Coexistence in a Fragmented Ecosystem* (pp. 209–232). Springer.

Giliba, R. A., Fust, P., Kiffner, C., & Loos, J. (2022). Multiple anthropogenic pressures challenge the effectiveness of protected areas in western Tanzania. *Conservation Science and Practice*, *4*(6). https://doi.org/10.1111/csp2.12684

Glamann, J., Hanspach, J., Abson, D. J., Collier, N., & Fischer, J. (2017). The intersection of food security and biodiversity conservation: A review. *Regional Environmental Change*, *17*(5), 1303–1313. https://doi.org/10.1007/s10113-015-0873-3

Gómez, C., White, J. C., & Wulder, M. A. (2016). Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *116*, 55–72. https://doi.org/10.1016/j.isprsjprs.2016.03.008

Hill, R., Miller, C., Newell, B., Dunlop, M., & Gordon, I. J. (2015). Why biodiversity declines as protected areas increase: The effect of the power of governance regimes on sustainable landscapes. *Sustainability Science*, *10*(2), 357–369. https://doi.org/10.1007/s11625-015-0288-6

Hoare, R. (2015). Lessons From 20 Years of Human–Elephant Conflict Mitigation in Africa. *Human Dimensions of Wildlife*, *20*(4), 289–295. https://doi.org/10.1080/10871209.2015.1005855

Janowicz, K., Gao, S., McKenzie, G., Hu, Y., & Bhaduri, B. (2020). GeoAI: Spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *International Journal of Geographical Information Science*, *34*(4), 625–636. https://doi.org/10.1080/13658816.2019.1684500

Jin, Z., Azzari, G., You, C., Di Tommaso, S., Aston, S., Burke, M., & Lobell, D. B. (2019). Smallholder maize area and yield mapping at national scales with Google Earth Engine. *Remote Sensing of Environment*, *228*, 115–128. https://doi.org/10.1016/j.rse.2019.04.016

Jóhannesdóttir, L., Alves, J. A., Gill, J. A., & Gunnarsson, T. G. (2017). Reconciling biodiversity conservation and agricultural expansion in the subarctic environment of Iceland. *Ecology and Society*, *22*(1). https://doi.org/10.5751/ES-08956-220116

John, E., Bunting, P., Hardy, A., Roberts, O., Giliba, R., & Silayo, D. S. (2020). Modelling the impact of climate change on Tanzanian forests. *Diversity and Distributions*, *26*(12), 1663–1686. https://doi.org/10.1111/ddi.13152

Jones, T., Bamford, A. J., Ferrol-Schulte, D., Hieronimo, P., McWilliam, N., & Rovero, F. (2012). Vanishing Wildlife Corridors and Options for Restoration: A Case Study from Tanzania. *Tropical Conservation Science*, *5*(4), 463–474. https://doi.org/10.1177/194008291200500405

Kays, R., Crofoot, M. C., Jetz, W., & Wikelski, M. (2015). Terrestrial animal tracking as an eye on life and planet. *Science*, *348*(6240), aaa2478–aaa2478. https://doi.org/10.1126/science.aaa2478

Kehoe, L., Romero-Muñoz, A., Polaina, E., Estes, L., Kreft, H., & Kuemmerle, T. (2017). Biodiversity at risk under future cropland expansion and intensification. *Nature Ecology and Evolution*, *1*(8), 1129–1135. https://doi.org/10.1038/s41559-017-0234-3

Kohi, E. M., de Boer, W. F., Peel, M. J. S., Slotow, R., van der Waal, C., Heitkönig, I. M. A., Skidmore, A., & Prins, H. H. T. (2011). African Elephants Loxodonta africana Amplify Browse Heterogeneity in African Savanna: Elephants Amplify Browse Heterogeneity. *Biotropica*, *43*(6), 711–721. https://doi.org/10.1111/j.1744-7429.2010.00724.x

Laso Bayas, J. C., Lesiv, M., Waldner, F., Schucknecht, A., Duerauer, M., See, L., Fritz, S., Fraisl, D., Moorthy, I., McCallum, I., Perger, C., Danylo, O., Defourny, P., Gallego, J., Gilliams, S., Akhtar, I. ul H., Baishya, S. J., Baruah, M., Bungnamei, K., … Wilson, J. W. (2017). A global reference database of crowdsourced cropland data collected using the Geo-Wiki platform. *Scientific Data*, *4*(1), 170136. https://doi.org/10.1038/sdata.2017.136

Laurance, W. F., Sayer, J., & Cassman, K. G. (2014). Agricultural expansion and its impacts on tropical nature. *Trends in Ecology and Evolution*, *29*(2), 107–116. https://doi.org/10.1016/j.tree.2013.12.001

Leite-Filho, A. T., Soares-Filho, B. S., Davis, J. L., Abrahão, G. M., & Börner, J. (2021). Deforestation reduces rainfall and agricultural revenues in the Brazilian Amazon. *Nature Communications*, *12*(1), 2591. https://doi.org/10.1038/s41467-021-22840-7

Lindenmayer, D., Cunningham, S., & Young, A. (2012). *Land use intensification: Effects on agriculture, biodiversity and ecological processes*. CSIRO publishing.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. *2008 Eighth IEEE International Conference on Data Mining*, 413–422.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-Based Anomaly Detection. *ACM Transactions on Knowledge Discovery from Data*, *6*(1), 1–39. https://doi.org/10.1145/2133360.2133363

Lohay, G. G., Weathers, T. C., Estes, A. B., McGrath, B. C., & Cavener, D. R. (2020). Genetic connectivity and population structure of African savanna elephants (*Loxodonta africana*) in Tanzania. *Ecology and Evolution*, *10*(20), 11069–11089. https://doi.org/10.1002/ece3.6728

Luhunga, P. M., Kijazi, A. L., Chang'a, L., Kondowe, A., Ng'ongolo, H., & Mtongori, H. (2018). Climate Change Projections for Tanzania Based on High-Resolution Regional Climate Models From the Coordinated Regional Climate Downscaling Experiment (CORDEX)-Africa. *Frontiers in Environmental Science*, *6*, 122. https://doi.org/10.3389/fenvs.2018.00122

Martin, E. H., Jensen, R. R., Hardin, P. J., Kisingo, A. W., Shoo, R. A., & Eustace, A. (2019). Assessing changes in Tanzania's Kwakuchinja Wildlife Corridor using multitemporal satellite imagery and open source tools. *Applied Geography*, *110*, 102051. https://doi.org/10.1016/j.apgeog.2019.102051

Metzger, K. L., Sinclair, A. R., Macfarlane, A., Coughenour, M., & Ding, J. (2015). Scales of change in the Greater Serengeti ecosystem. *Serengeti IV: Sustaining Biodiversity in a Coupled Human–Natural System'.(Eds ARE Sinclair, KL Metzger, JM Fryxell and SAR Mduma.) Pp*, 33–71.

Mokany, K., Ferrier, S., Harwood, T. D., Ware, C., Di Marco, M., Grantham, H. S., Venter, O., Hoskins, A. J., & Watson, J. E. M. (2020). Reconciling global priorities for conserving biodiversity habitat. *Proceedings of the National Academy of Sciences*, *117*(18), 9906–9911. https://doi.org/10.1073/pnas.1918373117

Nkwabi, A. K., Bukombe, J., Maliti, H., Liseki, S., Lesio, N., & Kija, H. (2018). An overview of biodiversity in tanzania and conservation efforts. *Global Biodiversity*, 295–340.

Ntukey, L. T., Munishi, L. K., Kohi, E., & Treydte, A. C. (2022). Land Use/Cover Change Reduces Elephant Habitat Suitability in the Wami Mbiki–Saadani Wildlife Corridor, Tanzania. *Land*, *11*(2), 307. https://doi.org/10.3390/land11020307

Olofsson, P., Stehman, S. V., Woodcock, C. E., Sulla-Menashe, D., Sibley, A. M., Newell, J. D., Friedl, M. A., & Herold, M. (2012). A global land-cover validation data set, part I: Fundamental design principles. *International Journal of Remote Sensing*, *33*(18), 5768–5788. https://doi.org/10.1080/01431161.2012.674230

Pettorelli, N., Wegmann, M., Skidmore, A., Mücher, S., Dawson, T. P., Fernandez, M., Lucas, R., Schaepman, M. E., Wang, T., O'Connor, B., Jongman, R. H. G., Kempeneers, P., Sonnenschein, R., Leidner, A. K., Böhm, M., He, K. S., Nagendra, H., Dubois, G., Fatoyinbo, T., … Geller, G. N. (2016). Framing the concept of satellite remote sensing essential biodiversity variables: Challenges and future directions. *Remote Sensing in Ecology and Conservation*, *2*(3), 122–131. https://doi.org/10.1002/rse2.15

Phalan, B., Onial, M., Balmford, A., & Green, R. E. (2011). Reconciling food production and biodiversity conservation: Land sharing and land sparing compared. *Science*, *333*(6047), 1289–1291. https://doi.org/10.1126/science.1208742

Planet Team. (2017). Planet application program interface: In space for life on Earth. *San Francisco, CA, 2017*, 40.

Robson, A. S., Trimble, M. J., Purdon, A., Young-Overton, K. D., Pimm, S. L., & van Aarde, R. J. (2017). Savanna elephant numbers are only a quarter of their expected values. *PLOS ONE*, *12*(4), e0175942. https://doi.org/10.1371/journal.pone.0175942

Rowhani, P., Lobell, D. B., Linderman, M., & Ramankutty, N. (2011). Climate variability and crop production in Tanzania. *Agricultural and Forest Meteorology*, *151*(4), 449–460. https://doi.org/10.1016/j.agrformet.2010.12.002

Roy, D. P., Huang, H., Houborg, R., & Martins, V. S. (2021). A global analysis of the temporal availability of PlanetScope high spatial resolution multi-spectral imagery. *Remote Sensing of Environment*, *264*, 112586. https://doi.org/10.1016/j.rse.2021.112586

Sawaya, K. E., Olmanson, L. G., Heinert, N. J., Brezonik, P. L., & Bauer, M. E. (2003). Extending satellite remote sensing to local scales: Land and water resource monitoring using high-resolution imagery. *Remote Sensing of Environment*, *88*(1–2), 144–156.

Schmitt, M., Hughes, L. H., Qiu, C., & Zhu, X. X. (2019). SEN12MS – A CURATED DATASET OF GEOREFERENCED MULTI-SPECTRAL SENTINEL-1/2 IMAGERY FOR DEEP LEARNING AND DATA FUSION. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *IV-2/W7*, 153–160. https://doi.org/10.5194/isprs-annals-IV-2-W7-153-2019

Schüßler, D., Lee, P. C., & Stadtmann, R. (2018). Analyzing land use change to identify migration corridors of African elephants (Loxodonta africana) in the Kenyan-Tanzanian borderlands. *Landscape Ecology*, *33*(12), 2121–2136. https://doi.org/10.1007/s10980-018-0728-7

Searchinger, T. D., Estes, L., Thornton, P. K., Beringer, T., Notenbaert, A., Rubenstein, D., Heimlich, R., Licker, R., & Herrero, M. (2015). High carbon and biodiversity costs from converting Africa's wet savannahs to cropland. *Nature Climate Change*, *5*(5), 481–486. https://doi.org/10.1038/nclimate2584

Shackelford, G. E., Steward, P. R., German, R. N., Sait, S. M., & Benton, T. G. (2015). Conservation planning in agricultural landscapes: Hotspots of conflict between agriculture and nature. *Diversity and Distributions*, *21*(3), 357–367. https://doi.org/10.1111/ddi.12291

Shaffer, L. J., Khadka, K. K., Van Den Hoek, J., & Naithani, K. J. (2019). Human-Elephant Conflict: A Review of Current Management Strategies and Future Directions. *Frontiers in Ecology and Evolution*, *6*, 235. https://doi.org/10.3389/fevo.2018.00235

Song, L., & Estes, L. (2023). ITSDM: Isolation forest-based presence-only species distribution modelling and explanation in R. *Methods in Ecology and Evolution*, 2041-210X.14067. https://doi.org/10.1111/2041-210X.14067

Song, X.-P., Hansen, M. C., Stehman, S. V., Potapov, P. V., Tyukavina, A., Vermote, E. F., & Townshend, J. R. (2018). Global land change from 1982 to 2016. *Nature*, *560*(7720), 639–643. https://doi.org/10.1038/s41586-018-0411-9

Strahler, A. H., Boschetti, L., Foody, G. M., Friedl, M. A., Hansen, M. C., Herold, M., Mayaux, P., Morisette, J. T., Stehman, S. V., & Woodcock, C. E. (2006). Global land cover validation: Recommendations for evaluation and accuracy assessment of global land cover maps. *European Communities, Luxembourg*, *51*(4), 1–60.

Suleiman, R. (2018). *Local and regional variations in conditions for agriculture and food security in Tanzania: A review*. https://doi.org/10.13140/RG.2.2.13136.76801

Tanzania National Bureau of Statistics. (2021). *Tanzania in Figures 2021*. Tanzania National Bureau of Statistics. https://www.nbs.go.tz/index.php/en/tanzania-in-figures/784-tanzania-in-figures-2021

Tilman, D., Clark, M., Williams, D. R., Kimmel, K., Polasky, S., & Packer, C. (2017). Future threats to biodiversity and pathways to their prevention. *Nature*, *546*(7656), 73–81. https://doi.org/10.1038/nature22900

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., & others. (2012). GMES Sentinel-1 mission. *Remote Sensing of Environment*, *120*, 9–24.

Tshipa, A., Valls-Fox, H., Fritz, H., Collins, K., Sebele, L., Mundy, P., & Chamaillé-Jammes, S. (2017). Partial migration links local surface-water management to large-scale elephant conservation in the world's largest transfrontier conservation area. *Biological Conservation*, *215*, 46–50. https://doi.org/10.1016/j.biocon.2017.09.003

Turner, M. G., & Gardner, R. H. (2015). *Landscape ecology in theory and practice: Pattern and process* (Second edition). Springer.

van Ittersum, M. K., van Bussel, L. G. J., Wolf, J., Grassini, P., van Wart, J., Guilpart, N., Claessens, L., de Groot, H., Wiebe, K., Mason-D'Croz, D., Yang, H., Boogaard, H., van Oort, P. A. J., van Loon, M. P., Saito, K., Adimo, O., Adjei-Nsiah, S., Agali, A., Bala, A., … Cassman, K. G. (2016). Can sub-Saharan Africa feed itself? *Proceedings of the National Academy of Sciences*, *113*(52), 14964–14969. https://doi.org/10.1073/pnas.1610359113

Vries, W. De, Vermeulen, S. J., Herrero, M., Carlson, K. M., Jonell, M., & Troell, M. (2018). Options for keeping the food system within environmental limits. *Nature*, *562*(7728), 519. https://doi.org/10.1038/s41586-018-0594-0

Wall, J., Wittemyer, G., Klinkenberg, B., LeMay, V., Blake, S., Strindberg, S., Henley, M., Vollrath, F., Maisels, F., Ferwerda, J., & Douglas-Hamilton, I. (2021). Human

footprint and protected areas shape elephant range across Africa. *Current Biology*, *31*(11), 2437-2445.e4. https://doi.org/10.1016/j.cub.2021.03.042

Xu, Y., Yu, L., Feng, D., Peng, D., Li, C., Huang, X., Lu, H., & Gong, P. (2019). Comparisons of three recent moderate resolution African land cover datasets: CGLS-LC100, ESA-S2-LC20, and FROM-GLC-Africa30. *International Journal of Remote Sensing*, *40*(16), 6185–6202.

Zacarias, D., & Loyola, R. (2018). Distribution modelling and multi-scale landscape connectivity highlight important areas for the conservation of savannah elephants. *Biological Conservation*, *224*, 1–8. https://doi.org/10.1016/j.biocon.2018.05.014

Chapter 1

A super-ensemble approach to map land cover types with high resolution over data-sparse African savanna landscapes

Lei Song [a*], Anna Bond Estes [b, c], Lyndon Despard Estes [a]

[a] Graduate School of Geography, Clark University, Worcester, MA, USA

[b] Environmental Studies Department, Carleton College, Northfield, MN, USA

[c] The Nelson Mandela African Institution of Science and Technology, Arusha, Tanzania

[*]Corresponding author: Lei Song

Published as:

Song, L., Estes, A. B., & Estes, L. D. (2023). A super-ensemble approach to map land cover types with high resolution over data-sparse African savanna landscapes. *International Journal of Applied Earth Observation and Geoinformation*, 116, 103152.

1.1 Abstract

Accurate and timely land cover products are critical inputs for landscape planning, and provide key information for biodiversity conservation and food security. However, poor mapping quality and low resolution are considerable issues in existing land cover maps over the African savanna, where land use is complex and changing rapidly, and necessary ground-truth data are sparse and hard to obtain.

To overcome this problem, to make optimal use of existing maps, and to minimize manual training data collection, we developed a three-stage ensemble method to make land cover maps. In the first stage, we extracted the consensus of multiple existing land cover products to generate fragmented pixel-wise training labels. In the second stage, we translated pixel-wise training labels to image-wise labels using Random Forest (RF) as a "gap-filling model", with temporal features extracted from Sentinel-1 time series, raw bands, and vegetation indices derived from PlanetScope basemaps. These image-wise labels were scored and edited by humans and the quality information was used in the next stage. For stage three, we trained a U-Net network based upon these image-wise labels, using Sentinel-1 time series and raw bands of PlanetScope basemaps as image features. Using the information on label quality, a quality-weighted loss function was used in the network to reduce the impact of noise in the training labels.

Using Northern Tanzania as a case study, the results demonstrate that ensembles of existing land cover maps provide a useful source of data for developing improved land cover maps over hard-to-classify, data-sparse landscapes. The Random Forest "gap-filling model" had an overall accuracy of 80.26% on our independent test dataset with 7 classes. The final U-Net model had an overall accuracy of 83.57%. This approach can be readily applied to other regions and extents (e.g., regional, global) and other data sources (e.g., Sentinel-2).

Keywords: Land cover classification, U-Net, Random Forest, African savanna, PlanetScope, Sentinel-1

1.2 Introduction

Timely and accurate land cover maps covering large extents are critical for many environmental applications, such as natural resources management, biodiversity conservation, and food security assessment (Anderson et al., 2017; Jin et al., 2019; Leite-Filho et al., 2021; Pettorelli et al., 2016; Song et al., 2018). However, current land cover maps are often inadequate for agricultural and ecological applications due to their limited spatial and temporal coverage and low accuracy (Gómez et al., 2016), particularly over data-sparse regions, such as Africa savanna. In the past two decades, satellite imagery with high spatial and temporal resolution has steadily opened new avenues for timely mapping of land cover over large areas (Cheng et al., 2020; Tong et al., 2020). Meanwhile, rapid improvements in machine learning techniques have led to dramatic gains in the accuracy of land cover classification (Campos-Taberner et al., 2020). Nevertheless, despite these gains, land cover mapping is still a major challenge in Africa's complex savanna landscapes with varying degrees of vegetation cover in space and time (Solbrig, 1996). One possible reason is the technical challenge of separating the woody and herbaceous components (Whitley et al., 2017). For these areas, the available global land cover products (Buchhorn et al., 2020a, 2020b; Congalton et al., 2017; Xu et al., 2019) that have high error rates when applied at regional scales, and perform poorly in savanna environments. Furthermore, existing savanna monitoring studies generally use coarse to medium resolution imagery that do not effectively represent fine grained components, such as nonforest trees (Abdi et al., 2022) and fields in smallholder agricultural systems (Jin et al., 2019; Kerner et al., 2020).

This latter problem of image resolution is being increasingly overcome by the mission of new satellites. Since late 2017, PlanetScope sensors started to supply near-daily satellite

imagery with 3.7 m spatial resolution, which increases the possibility of obtaining clear observations during the rainy season (Planet Team, 2017; Roy et al., 2021). One of the higher-level derived products from this daily imagery are surface reflectance basemaps, which provide monthly to biannual composites of daily imagery in an analysis-ready format, substantially reducing the amount of pre-processing work (Estes et al., 2022) that must be undertaken to develop cloud-free mosaics. The recent Norway International Climate and Forests Initiative Imagery Program (NICFI) makes basemaps collected over tropical regions free to the public for sustainability-focused, non-commercial research (Norway's International Climate and Forest Initiative (NICFI), 2020), which is a game-changer for tropical land cover and land use monitoring. The availability of Sentinel-1 imagery also improves savanna monitoring because it is not affected by cloud cover (Torres et al., 2012) that frequently obstructs optical sensors over much of the tropical savanna biome (Roy et al., 2021). Given the recent release of the NICFI dataset, there are still relatively few studies that have applied these data for large area land cover mapping, particularly in combination with Sentinel-1 data (Vizzari, 2022). Existing applications of NICFI basemaps typically help visual image interpretation (Pascual et al., 2022; Rienow et al., 2022; Sugimoto et al., 2022) or evaluate the effectiveness of this dataset for delineating land cover (Aquino et al., 2022; Awuah and Aplin, 2021; Vizzari, 2022) within relatively small areas (< 100000 km^2), although efforts at larger sub-national (e.g. cropland mapping, Rufin et al., 2022) to continental extents are beginning to emerge (e.g. tree cover mapping; Reiner et al., 2022).

Alongside the newly available sources of high-frequency, high-resolution satellite imagery, there have been corresponding increases in the computational power needed to process large datasets, while advances in deep learning (DL) models have led to dramatic improvements

in land cover mapping. The models based on the fully convolutional network (FCN) are among the most widely used deep learning architectures for land cover mapping (Chamorro Martinez et al., 2021; Solórzano et al., 2021; Volpi and Tuia, 2016) because they can achieve pixel-wise segmentation (Long et al., 2015).

A key obstacle to the wider adoption of these models for land cover mapping is the need to collect large, task-specific training and reference datasets. Despite recent investments in developing global ground truth datasets, which include observations within tropical savannas (Burke and Lobell, 2017; Laso Bayas et al., 2017; Schmitt et al., 2019), as well as new strategies to extend training samples (e.g., transfer learning and data augmentation; Shorten and Khoshgoftaar, 2019; Torrey and Shavlik, 2010), the available data are typically insufficient for training deep learning models over large areas. Since collecting labelled training data is time and labor-intensive, a more efficient labeling strategy is needed to meet the need for timely high-resolution land use/land cover products in these data-sparse but highly dynamic landscapes.

To overcome the challenge of developing the large label datasets that are needed to train deep learning models that can improve the ability to map savannas environments, our study had two primary objectives. The first was to develop a more automated and objective approach for generating labels that minimized the amount of manual effort. The second was to design a modeling approach that can account for and minimize the impact of label error (following Elmes et al., 2020).

To satisfy the first objective, we developed a technique for creating synthetic labels, in which several existing land cover products (Buchhorn et al., 2020a, 2020b; Congalton et al., 2017; Xu et al., 2019) were combined into a consensus land cover map, providing an initial set of labels that were more reliable than any of the individual inputs (following the rationale of Fritz et

al., 2011), but provided only partial coverage of the study region (areas lacking consensus were excluded). To fill the missing data in the resulting synthetic labels, we trained a Random Forest (RF; Breiman, 2001) model to fill the gaps in selected tiles, with predictors based on temporal features extracted from Sentinel-1 time series, as well as the raw bands of the PlanetScope basemaps and additional derived vegetation indices. We then visually assessed and manually edited the resulting gap-filled labels, producing a final set of fully labelled tiles, along with a quantitative measure of label quality. To meet the second objective, we used the resulting tiles to train a U-Net land cover model using a label quality-weighted loss function to minimize the impacts of label error.

We then applied this model to map the complex savanna landscapes of a 243,416 km^2 region in Northern Tanzania, in the process demonstrating an efficient, semi-automated approach for training a model capable of generating improved, high resolution land cover data in hard-to-map, data sparse environment.

1.3 Materials and methods

1.3.1 Study area

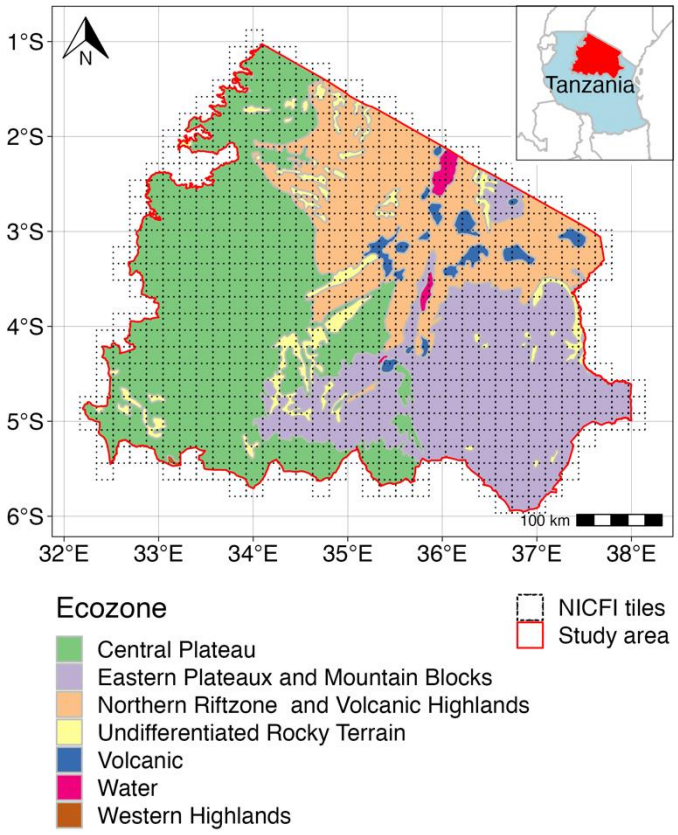

Figure 1-1. The ecozones of the study area overlaid by the PlanetScope basemap tiling grid

We applied our proposed method to northern Tanzania (Figure 1-1). This region is ecologically complex, comprising more than 4 ecozones, where the climatic conditions, soil moisture, and farming systems are different (Figure 1-1) (Sebastian, 2009). Remote-sensing-based mapping of savanna in East Africa is challenging because savanna is a heterogeneous landscape (Solbrig, 1996) with varying degrees of vegetation cover and spectral similarities among land cover types (Tsalyuk et al., 2017; Zhang et al., 2019). These characteristics make this area an ideal case study, as its complex land cover provides an important baseline for our

method, and the environmental context means that there is an urgent need for accurate land cover maps.

1.3.2 Datasets

1.3.2.1 Satellite imagery

In this study, we queried 543 quads of PlanetScope Tropical Normalized Analytic Biannual basemap from 2017 to 2018, which are provided by NICFI program. We collected imagery covering two seasons in each quad: December 2017 – May 2018 (season 1), and June 2018 – November 2018 (season 2). Having the spectral signature from multiple seasons helps to differentiate land cover types (Estes et al., 2022). To incorporate temporal features, we also used harmonic regression coefficients of Sentinel-1 time series (2017-10-01 – 2018-09-30). Harmonic regression coefficients summarize critical temporal features, enhancing the ability to differentiate land cover types based on seasonal information contained in the series (Moody and Johnson, 2001), while significantly reducing the number of raw images used. The coefficients were fitted on level-1 Ground Range Detected (GRD) Interferometric Wide Swath (IW) images acquired with dual polarization (VV + VH). Taking tile 1227-1002 as an example, Figure 1-2 shows two seasonal NICFI basemaps, harmonic coefficients of Sentinel-1 dB in VV and VH polarization, and the spectral and temporal signature of land cover samples. The complete details of the images and image preprocessing can be found in section 4.5A.1 and 4.5A.2 of 4.5Appendix A.

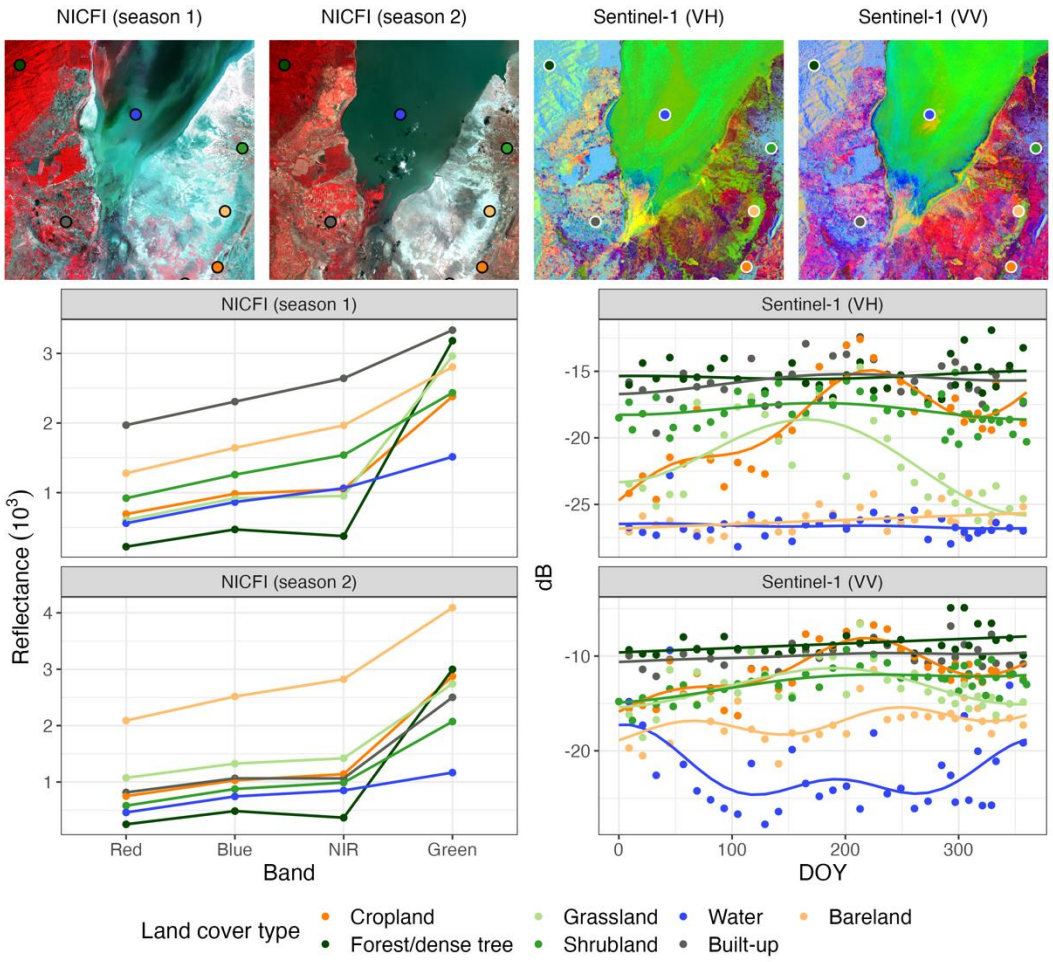

Figure 1-2. Seasonal NICFI basemap in false-color, harmonic coefficients (RGB: Slope, $\cos\left(\frac{2\pi t}{d_{yr}}\right)$, Intercept) of Sentinel-1 dB in VV and VH polarization, spectral signature, and temporal signature of land cover samples in an example tile (1227-1002)

1.3.2.2 Pixel-level land cover (LC) reference labels

We developed an independent validation (n = 1286) dataset to evaluate the quality of land cover products used in our study (section 1.3.2.3), and to assess the performance of the land cover model. Two 10m × 10m squares in each quad within the study area were randomly selected, and hand-labeled based on visual interpretation based on several virtual globe basemaps (e.g. Bing or Google Maps), satellite imagery in this study (section 1.3.2.1), prior knowledge,

and web searches of local landscape pictures. Despite the small size of the squares, they occasionally contained more than one class. In these cases, we shifted the squares so that they covered a single class.

1.3.2.3 Land cover (LC) products

Table 1-1. Land cover products used and their accuracies over the study region.

Product name	Product type	Year	Resolution	Overall accuracy
CGLS_LC100m	Land cover	2018	100 m	67.70%
FROM-GLC 2017v1	Land cover	2017	30 m	60.13%
GFSAD30	Crop mask	2015	30 m	82.93%
TanSIS	Crop mask	2018	250 m	82.39%

To initialize the training dataset for our land cover mapping models (see section 1.3.3), we prepared an ensemble of existing land cover products, which we turned into consensus labels. After pre-assessment (see section 4.5A.3), we selected 4 products to use (Table 1-1): Copernicus global land cover map (CGLS_LC) (Buchhorn et al., 2020b), Finer Resolution Observation and Monitoring – Global Land Cover (FROM-GLC) (Gong et al., 2019), Global Food Security-support Analysis Data (GFSAD) cropland extent of Africa (Congalton et al., 2017), and a cropland layer produced by the Tanzania Soil Information Service (TanSIS) (Walsh et al., 2018). CGLS_LC has 23 classes according to UN-FAO's Land Cover Classification System (LCCS). FROM-GLC includes 10 main land cover classes. GFSAD and TanSIS are cropland masks, although GFSAD has an extra water class. Even though each product was separately validated, the quality of these products in our study area is unclear. We, therefore, used the independent validation dataset (section 1.3.2.2) to evaluate their accuracies, which are listed in Table 1-1 (complete details of the assessment are in Table A-2). Besides gridded LC products,

OpenStreetMap (OSM) vector layers also were used as ancillary datasets in both generating consensus labels and land cover classification.

1.3.3 Modeling approach

The land cover mapping approach we developed has three main parts (Figure 1-3). In the first part, we created an ensemble map from the selected land cover products (section 1.3.2.3) as land cover reference labels. In the second part, we trained a pixel-based Random Forest model, selected a group of tiles, and used the model to fill the gaps between labelled fragments obtained in part 1 in these tiles, in order to make complete image-wise labels. A deep learning network was then trained using these gap-filled labels in the third part.

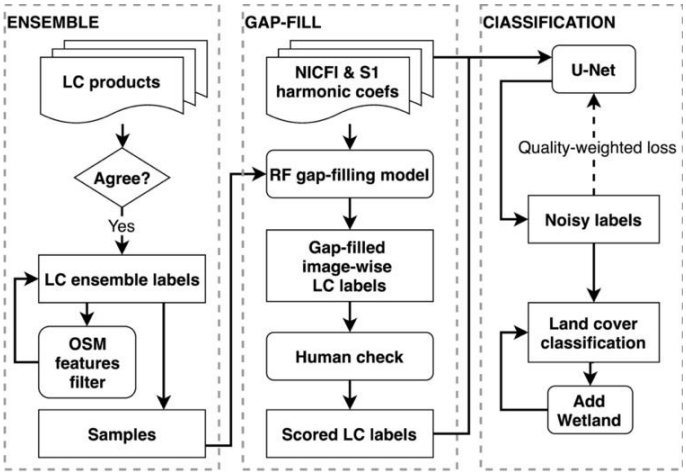

Figure 1-3. The workflow of the proposed approach (OSM stands for OpenStreetMap)

1.3.3.1 Part 1: Ensemble multiple land cover (LC) products

Several previous studies have demonstrated the effectiveness of deriving training labels from existing land cover products (Ren et al., 2022; Yang and Huang, 2021; Zhang et al., 2021). We made consensus land cover labels from assembled land cover products (Table 1-1). These

land cover products were made in different years based on varied features, therefore identifying where these products agree helps to increase confidence in the accuracy of the underlying land cover classification (Fritz et al., 2010; Pérez-Hoyos et al., 2020), while also providing information on how stable and how difficult a landscape is to classify. We assume that pixels where all maps agree represent stable landscapes, while the pixels in which maps disagree are either areas undergoing rapid changes (e.g., fallows) or those that are hard to distinguish from other cover types (e.g., degraded savannas versus croplands).

To create the ensembles, we aggregated the land cover types into 8 common classes that could be extracted from the different taxonomies of these land cover products, which were cropland, forest/dense tree, shrubland, grassland, water, wetland, built-up, and bareland (which in this region are typically degraded savannas) (Doggart et al., 2020). The selected land cover products were either multi-class land cover products or binary cropland layers. Multi-class land cover maps were first reclassified into the same classes (Table A-1), combined, and then the consensus areas were extracted from these layers. Because cropland is a difficult class to predict and thus was assigned poor quality consensus labels, a different strategy was applied to the cropland class. GFSAD30 and TanSIS were combined together to mask out cropland areas to make consensus labels of other classes. GFSAD30 was then used singly as the cropland label at this stage due to its higher spatial resolution and classification accuracy.

The resulting consensus labels still contained significant noise because the input maps were derived from low to moderate spatial resolution imagery (Table 1-1). These errors were mainly concentrated along the boundaries of different cover types, such as roads, rivers, and built-up regions. To reduce the impact of boundary uncertainties, we used buffered OSM layers, including roads, rivers, and buildings, to mask out such areas. This process resulted in a set of

consensus land cover labels that were highly fragmented (Figure 1-4A) and were thus not optimal for training fully convolutional neural networks, for which image-based training labels are optimal for providing the model with spatial context.

1.3.3.2 Part 2: Creating gap-filled land cover labels

To convert the fragmented consensus labels into image-wise labels (Figure 1-4B) suitable for training a U-Net, we trained a pixel-based Random Forest (Breiman, 2001) model using a random sample of consensus LC labels. As the moderate-resolution land cover maps miss many small inland waterbodies and have low location precision in settlements, we removed waterbodies and built-up areas from the consensus labels and instead sampled labels for these two classes from the OSM layers. We did not explicitly model wetlands, because they constitute a very rare class without unique features in our study area. To develop the model, we analyzed variable importance and selected the best hyper-parameters (see details in Figure A-4). The model was trained using the 4 bands of semi-annual NICFI basemaps, and the harmonic regression coefficients extracted from both polarizations of the Sentinel-1 time series (see section 1.3.2.1). Additional features included the normalized difference vegetation index (NDVI), soil-adjusted vegetation index (SAVI), two-band enhanced vegetation index (EVI), and atmospherically resistant vegetation index (ARVI) (Huete, 1988; Jiang et al., 2008; Jin et al., 2019; Kaufman and Tanre, 1992; Tucker, 1979) calculated from each basemap pair. To minimize the computation costs of training the neural network, we split the original quad size of 4096 × 4096 pixels into 8 × 8 sub tiles of 512 × 512 pixels (Figure 1-4). After training the gap-filling model, 4 sub tiles within each quad were randomly selected to generate predicted labels, with 3 labels reserved for training the neural network, and one reserved for validation.

The resulting gap-filled land cover labels were imperfect (see Figure 1-4B), therefore in a second step, we manually checked and edited labels where needed. Final label quality was assessed against two dimensions: the correctness of the label after editing and the difficulty of the tile to be classified. The correctness of the label was assessed based on visual interpretation of the underlying land cover in the basemap imagery, while the difficulty of labeling was graded based on how many edits were made to the predicted label tile, with more edits indicating increasing difficulty. Both measures were graded from low (1) to high (5). We refer to the resulting scored and edited labels as human refined labels (Figure 1-4C), which, along with their quality information, were used to train the neural network.

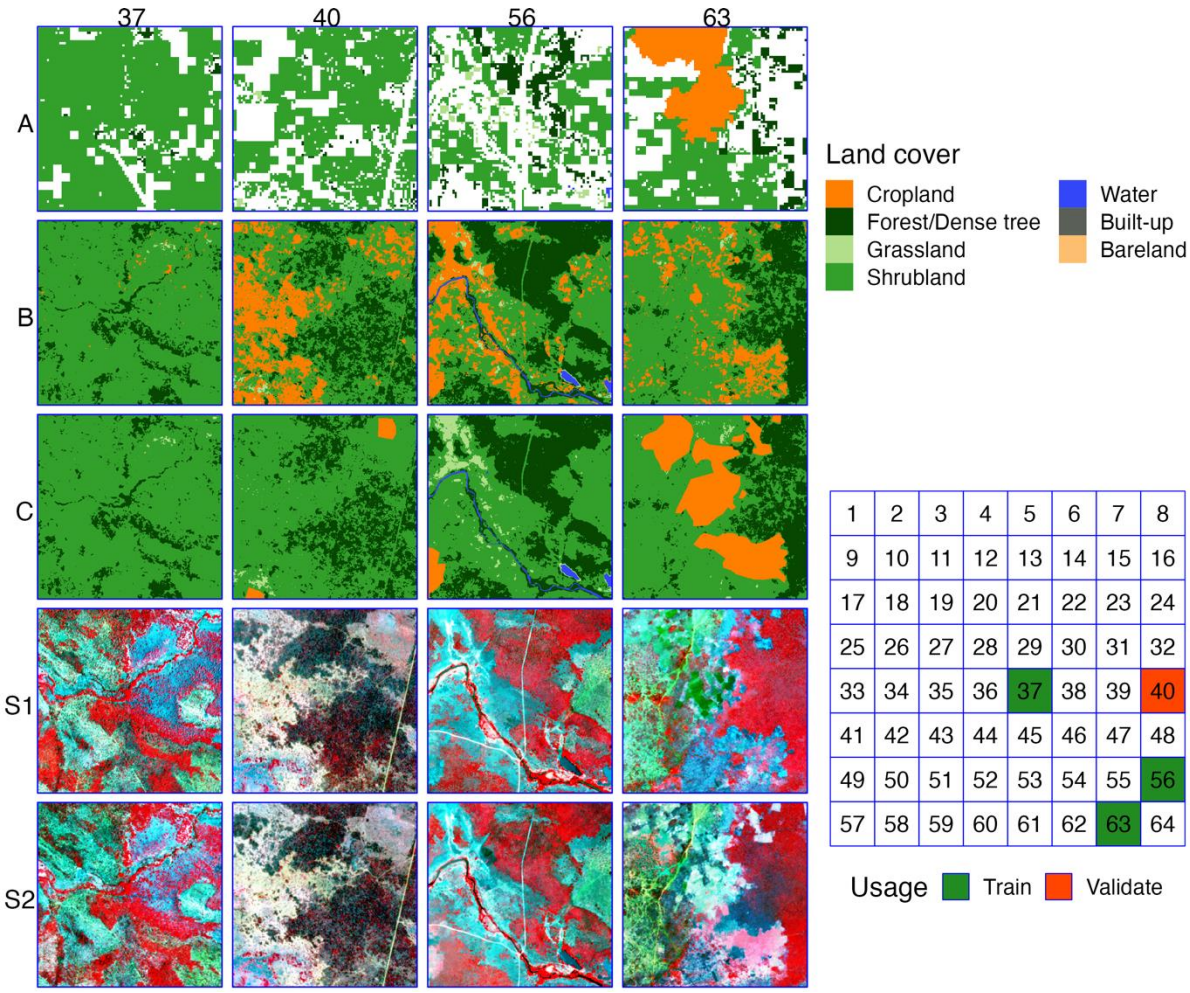

Figure 1-4. Sub-tiling system and different types of label examples (A are the weak ensemble labels, B are the gap-filled labels produced by a Random Forests model, C are the human refined labels, and S1 and S2 are false-color composites of PlanetScope NICFI basemaps in season 1 and season 2)

1.3.3.3 Part 3: Land cover classification with U-Net

We applied a widely used deep learning architecture, U-Net (section 4.5A.4 and Figure A-2), to train the image-based land cover model. Compared to more complex architectures (e.g., Deeplab), U-Net requires fewer training samples and achieves good performance with a substantially lower computational cost (Ronneberger et al., 2015). Thus, U-Net is a popular

method for land cover classification, particularly when training examples are limited (Rakhlin et al., 2018).

Optimization algorithms play an important role in training a deep neural network (Sun, 2020). A good optimizer enables the network to obtain the optimal weight matrix efficiently. In this study, we selected a recently proposed combined optimization method (AdaBound) that works like adaptive methods at the early stage of training to get fast training speeds, then smoothly transforms to Stochastic Gradient Descent (SGD) at the end to attain good overall generalization and model performance (Luo et al., 2019).

The learning rate for gradient descent is also critical in neural network training that impacts the model's ability to converge on a solution (Bengio, 2012). In this paper, we applied a combined learning rate scheduler that included multiple constant and cyclical learning rates (Smith, 2017) for all experiments.

In a convolutional neural network, data augmentations, such as scale and rotation, are effective strategies to increase the variance of training data and improve the generalizability of the network (Perez and Wang, 2017; Shorten and Khoshgoftaar, 2019). At training time, we randomly flipped the original images horizontally or vertically. Diverse imaging conditions of CubeSats lead to photographic variations in scale and changes of illumination in PlanetScope imagery (Houborg and McCabe, 2018). These variations effectively mimic the brightness shift that is typically used for data augmentation; therefore, we did not implement this particular augmentation strategy. We did not employ rotations or stretching because they may alter image-level labels or break the spatial symmetry.

To reduce the impact of class imbalance and noisy labels, we weighted loss by class frequency as well as the quality of labels in each training image (Eq. (1-1). The dimensions of

the input for the loss function were (B, C, H, W), where B is the number of images in each mini-batch, C is the number of classes, H is the height of the image, and W is the width of the image. N is the overall number of pixels, which equals $B \times H \times W$.

$$loss(x, y) = \frac{\sum_{b=1}^{B} l_b \times w_{cb} \times w_{db}}{\sum_{n=1}^{N} w_{yn} \cdot 1\{y_n \neq index_{ignore}\}} \tag{1-1}$$

l_b is the sum of class-balanced cross entropy loss per pixel in an image, then:

$$l_b = \sum_{h=1}^{H} \sum_{w=1}^{W} -w_{y_{hw}} log \frac{exp(x_{hw,\, y_{hw}})}{\sum_{c=1}^{C} exp(x_{hw,c})} \cdot 1\{y_{hw} \neq index_{ignore}\} \tag{1-2}$$

where x is the input, y is the target, w is the class-balanced weight, w_c is the normalized weight of label correctness, and w_d is the normalized weight of label difficulty.

The original measure of label correctness c ranges from 1 (least correct) to 5 (most correct), which are then normalized to weights (0 – 1) using a logistic function $w_c = \frac{1}{1+e^{-k(c-c_m)}}$, where k is the growth rate and c_m is the c value of the midpoint. In w_c, we only changed k and used the constant value 2.5 for c_m, which is the median of original correctness values. The original value for labeling difficulty d ranges from 1 to 5 (least to most difficult), which are rescaled to a user defined range [1, s_{max}] as weights using equation $w_d = \frac{(s_{max}-1)\times(d-d_{min})}{(d_{max}-d_{min})} + 1$. The rescaling functions are flexible to be designed in any forms (e.g., linear) based on different modeling demands. We picked these functions based on their regularization ability and the quality condition of the labels in this project.

The correctness weight reduces the contribution of images with poor label quality to the calculation of the loss function, thereby decreasing the risk that the network learns incorrect information from noisy labels. In contrast to correctness weights, the role of difficulty weights is

to force the model to pay more attention to images that are hard to classify, which helps improve the generalizability of the model to other areas or time intervals. However, difficulty weights that are too high may cause the network to overfit on difficult labels, therefore we recommend smaller values of s_{max}, generally from 1 – 1.2. An analysis of weight parameters of quality-weighted and class-balanced loss was given in section 4.5A.6 of 4.5Appendix A. In our study, when growth rate k of the logistic function used to calculate the correctness weight (w_c) is set to 1.5, the best performance (AA = 0.880 ± 0.002) was achieved (Figure A-3). The values of s_{max}, used to calculate the difficulty weight (w_d), from 1 to 1.1 were reasonable choices for this study (section 4.5A.6, Figure A-3).

1.3.4 Accuracy assessment

To assess the LC products and modeling results, we used the true negative rate (TNR, also called specificity), negative predictive value (NPV), user's accuracy (UA, also called consumer's accuracy, precision, or positive predictive value), producer's accuracy (PA, also called recall, true positive rate or sensitivity), balanced accuracy (BA), and the F1 score (Barsi et al., 2018; Brodersen et al., 2010; Elmes et al., 2020; Olofsson et al., 2014). For U-Net training, we also used intersection over union (IoU) and the average (mIoU). The calculation details for these metrics are provided in section 4.5A.5 of 4.5Appendix A.

1.4 Results

1.4.1 Land cover label gap-filling

The Random Forests label gap-filling model (number of trees: 1000, number of independent variables: 28, mtry: 11) was trained with 1.24×10^6 samples randomly selected from the ensembled land cover products (roughly 48% of the area). Twenty percent of the samples were used for hyper-parameter tuning with a grid search and model validation (Table 1-2). The independent reference dataset (section 1.3.2.2) was used to assess model performance (Table 1-2).

Table 1-2. The evaluation of the random forest gap-filling model

Class	Validation PA	Independent test (overall accuracy: 80.26%)					
		TNR	NPV	UA	PA	BA	F1 score
Cropland	84.24%	88.17%	96.21%	80.05%	93.18%	90.67%	86.12%
Forest/Dense tree	95.04%	98.23%	98.56%	64.23%	69.02%	83.62%	66.54%
Grassland	87.52%	93.30%	94.99%	69.06%	75.26%	84.28%	72.03%
Shrubland	86.06%	94.47%	85.62%	88.28%	72.43%	83.45%	79.57%
Water	96.74%	99.57%	99.93%	89.14%	98.01%	98.79%	93.36%
Built-up	73.62%	99.46%	98.77%	83.70%	69.06%	84.26%	75.68%
Bareland	96.78%	99.82%	99.35%	81.13%	53.75%	76.79%	64.66%
Average	88.57%	96.14%	96.21%	79.37%	75.82%	85.98%	76.85%

The model selection process showed that radar backscatter (Figure A-4) was an important feature for land cover classification, with variables representing harmonic regression coefficients describing one intra-annual seasonal cycle being much more influential in distinguishing land cover types than those characterizing two seasonal cycles, while the coefficient of $cos\left(\frac{2\pi t}{d_{yr}}\right)$ was more important than $sin\left(\frac{2\pi t}{d_{yr}}\right)$, which is also evident in Figure 1-2.

Optical images captured in the growing season were more influential than those from the off-season (Figure 1-2), while among the spectral features, NIR and vegetation indices contributed more than the visible bands. The trained Random Forest had an overall accuracy of 80.26% and an average F1 of 76.85%, but with substantial performance differences across different land cover types (Table 1-2 & Figure A-5).

The resulting gap-filled models produced a set of 2,572 image-wise labels that were generated purely by machine intelligence, which were then improved using human supervision. Each label was manually checked and refined as needed, requiring an average of 0.8 minutes per label and a total estimated effort of 34 hours (see Table A-3 in 4.5Appendix A for further details on effort). Many (39%) labels had good quality and did not need editing, and thus had high correctness and low difficulty scores (Figure 1-5). In all, 61% of the labels were edited, with the majority of these (85%) needing only minor editing (difficulty score of 1-2), while the remaining 15% needed moderate to extensive editing (difficulties of 3-5). A small quantity (6%) of labels had correctness scores lower than five because these were too complex to be edited. The resulting human-refined, image-wise labels and their quality information were used to train the U-Net land cover segmentation model.

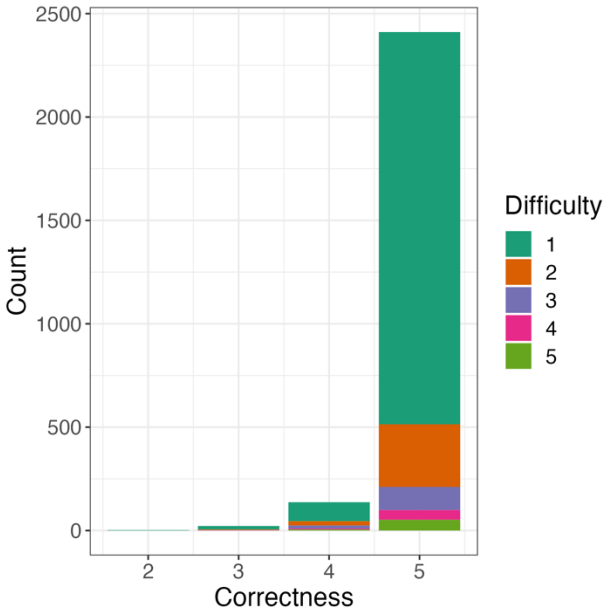

Figure 1-5. Distribution of human refined label quality

1.4.2 Land cover classification with U-Net

To train the model, we used original spectral bands of two seasonal NICFI images and intercept, slope, and coefficient of $cos\left(\frac{2\pi t}{d_{yr}}\right)$ of Sentinel-1 time series as input channels of U-Net, which were selected based on their variable importance values for Random Forests (Section 1.4.1 and Figure A-4). Other inputs were not used in order to minimize computational demand. For the learning rate scheduler, we selected values that ranged from 0.001 in earlier epochs to 0.0001 in the final epochs (Figure 1-6A). The selected validation sub tiles (section 1.3.3.2 and Figure 1-4) were used for model tuning and model validation, and the independent reference dataset (section 1.3.2.2) was also used for final model testing (Table 1-3). The computational environments are described in section 4.5A.4 in 4.5Appendix A.

Table 1-3. Validation, independent test accuracy and prediction confidence of U-Net model

Class	Validation PA[1]	Independent test (overall accuracy: 83.57%)						Prediction confidence (95% interval)
		TNR	NPV	UA	PA	BA	F1 score	
Cropland	85.20±0.78%	96.34%	95.84%	92.75%	91.80%	94.07%	92.27%	81.27±15.10%
Forest/Dense tree	97.47±1.14%	96.50%	98.89%	50.26%	76.47%	86.49%	60.65%	83.95±17.48%
Grassland	83.03±1.02%	94.02%	95.61%	72.25%	78.29%	86.16%	75.15%	82.82±17.12%
Shrubland	80.91±1.47%	93.48%	87.27%	87.07%	76.29%	84.89%	81.32%	82.56±16.68%
Water	97.73±0.20%	99.05%	99.93%	78.80%	98.01%	98.53%	87.36%	94.89±12.18%
Built-up	88.51±0.35%	99.93%	99.86%	98.17%	96.41%	98.17%	97.29%	70.11±18.41%
Bareland	96.22±0.19%	99.44%	99.84%	68.93%	88.75%	94.09%	77.60%	79.93±18.62%
Average	89.87±0.10%	96.97%	96.75%	78.32%	86.57%	91.77%	81.66%	82.43±16.53%

[1]Average PA of last 10 epochs

The model training curves (Figure 1-6) showed that the U-Net model achieved relatively high accuracy for almost every class after just a few epochs (10-25), with gradual improvement thereafter with fluctuations in the curve becoming more stable in later epochs. The mIoU curve reveals that the model was less effective in achieving shape accuracy than overall accuracy, given the slower rate of improvement and significant fluctuations. The learning curves for the cropland and built-up classes showed a characteristic pattern of rapid but steady growth until reaching a plateau, with high values for each class achieved on the independent accuracy assessment (92-97%, Table 1-3) test. In contrast, the training curves for the forest/dense tree and bareland classes rapidly jumped to a plateau, with larger fluctuations evident in the forest class.

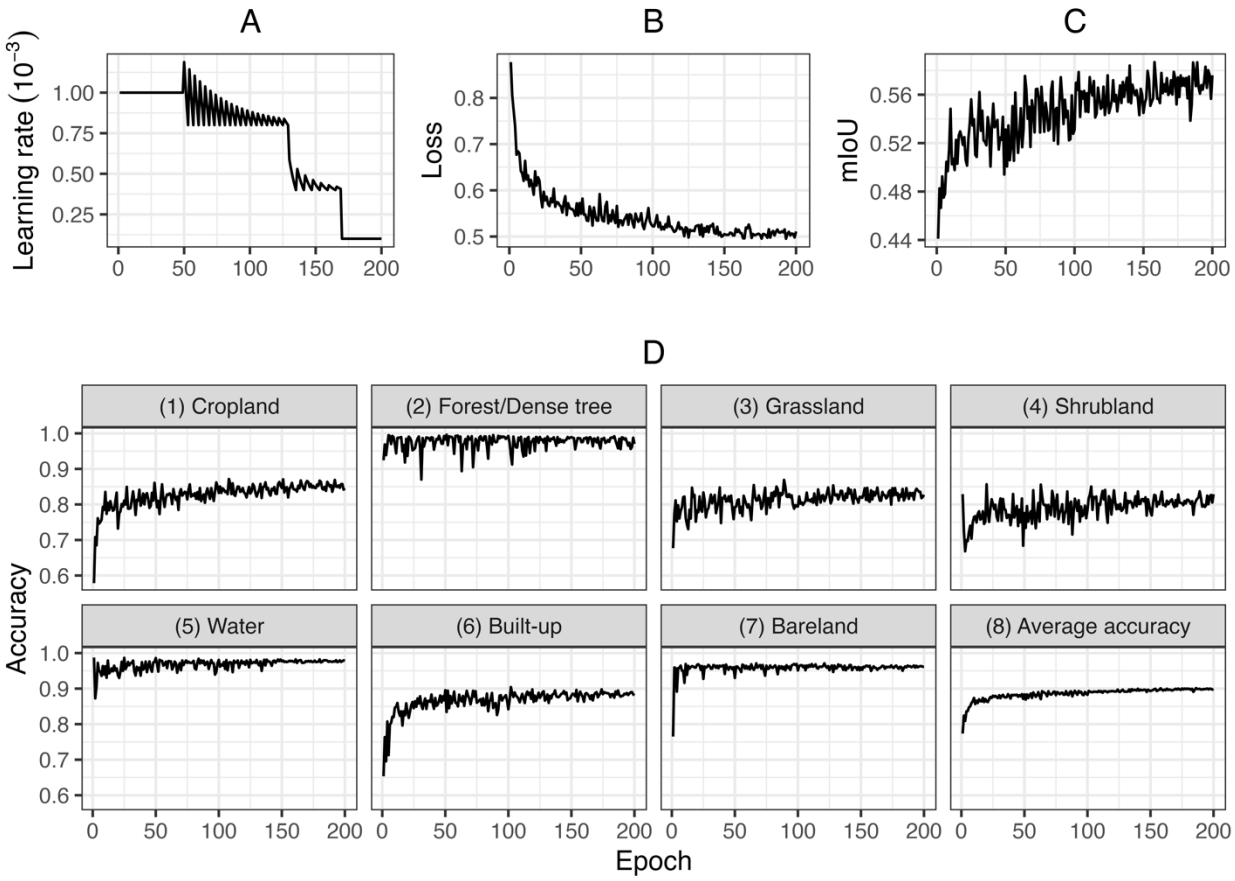

Figure 1-6. Evolution of the learning rate (A), loss function (B), mean IOU (C), accuracy of each class (D1-D7), and average accuracy (D8)) on the validation dataset during model training

This likely reflects the same tendency to over-fit these two classes that were evident in the Random Forest gap-filling model (section 1.4.1). The same behavior is also evident in the learning curve for water, but in this case, the model is able to effectively generalize for this class (98-99% accuracy; Table 1-3) because of its relatively unique and globally consistent spectral characteristics. Grassland and shrubland had learning curve progressions similar to those of the cropland and built-up classes, but with substantial fluctuations. These two types were more confusing for the network, given their spectral similarity with each other and with croplands.

The U-Net model achieved an overall accuracy of 83.57% and an average F1 of 81.66% on our independent test dataset. The average balanced accuracy across classes was 91.77% (range 84.89-98.53%), with average producer's accuracy (PA) of 86.57% (range 76.29-98.01%) and average user's accuracy (UA) of 78.33% (range 50.26-98.17%). Forest/dense trees had a high commission error (UA = 50.26%) and were often classified in places that were in fact shrublands (Figure 1-7), while also having a high commission error (UA = 87.07%) as it was often confused with grassland (Figure 1-7). Besides forests/dense trees, grassland and shrubland were the two hardest classes for U-Net to learn, particularly shrubland, which has diverse features that are easily confused with croplands, forests, and grasslands (Figure 1-7).

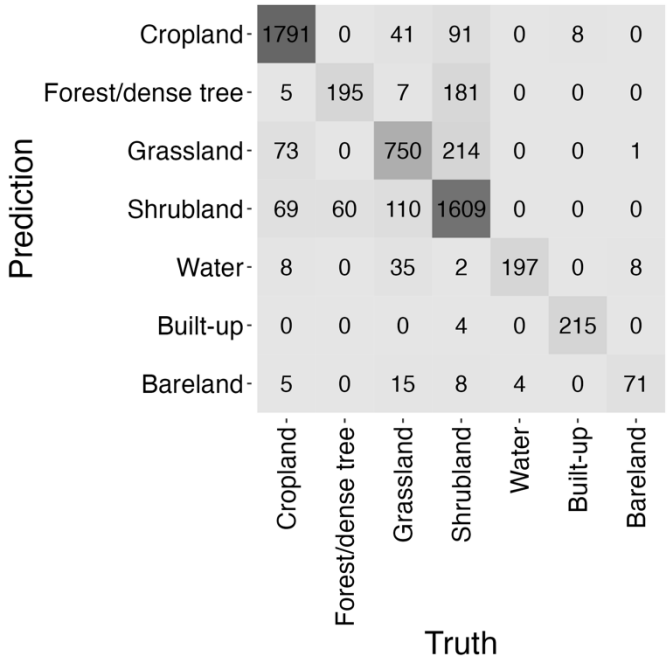

Figure 1-7. Confusion matrix heatmap of the independent test of U-Net classification model

Besides the land cover map, U-Net also produced a map of classification confidence (Figure 1-8) with values ranging from 18 to 99 to provide a reference for downstream studies such as yield estimation or land cover change analysis, following recommendations by Elmes et

al (2020). The classification confidence is the pixel-wise maximum value of class probabilities produced by softmax after U-Net (Figure A-2). Over the whole study area, the cropland, forest/dense trees, grassland, shrubland, and bareland classes had fairly similar average confidence scores (80-84%; Table 1-3), while water had the highest mean score (94.89%; Table 1-3) and built-up the lowest (70.11%; Table 1-3). It is noteworthy that there was some noise in the predicted water class due to dark objects in the original images, such as shadows.

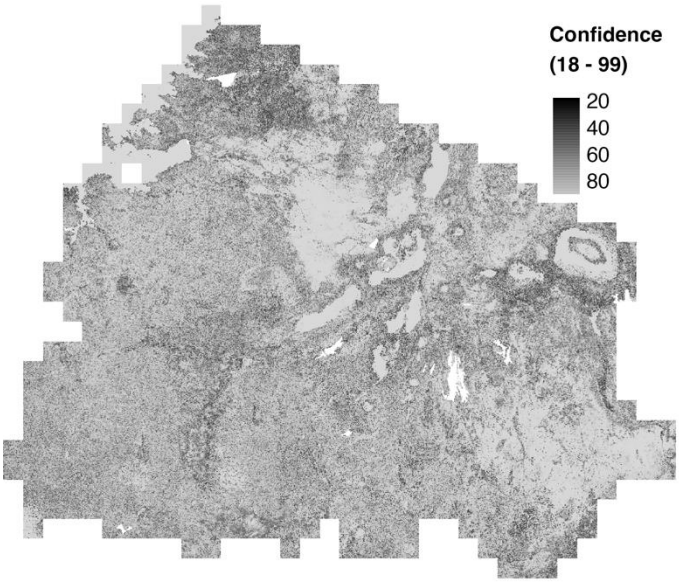

Figure 1-8. The confidence map of land cover classification by U-Net

1.4.3 Final land cover map

After completing the final land cover classification, we added the wetland class by rasterizing the OSM layer (Figure 1-9A). Because the wetland class was not calculated by U-Net, we did not assign this class any confidence value (Figure 1-8). Reasonably, the U-Net model has higher confidence over homogeneous areas, such as water bodies, grassland in the north, and shrubland in the southeast (Figure 1-8). The model has low confidence over the northwest area

(Figure 1-8), where it is a complex mixture of cultivated crops, tree crops, built-up, and native vegetation. Similar performance was attained over the foothills and hillside of Mount Kilimanjaro, where people are growing cultivated crops and coffee mixing with native vegetation.

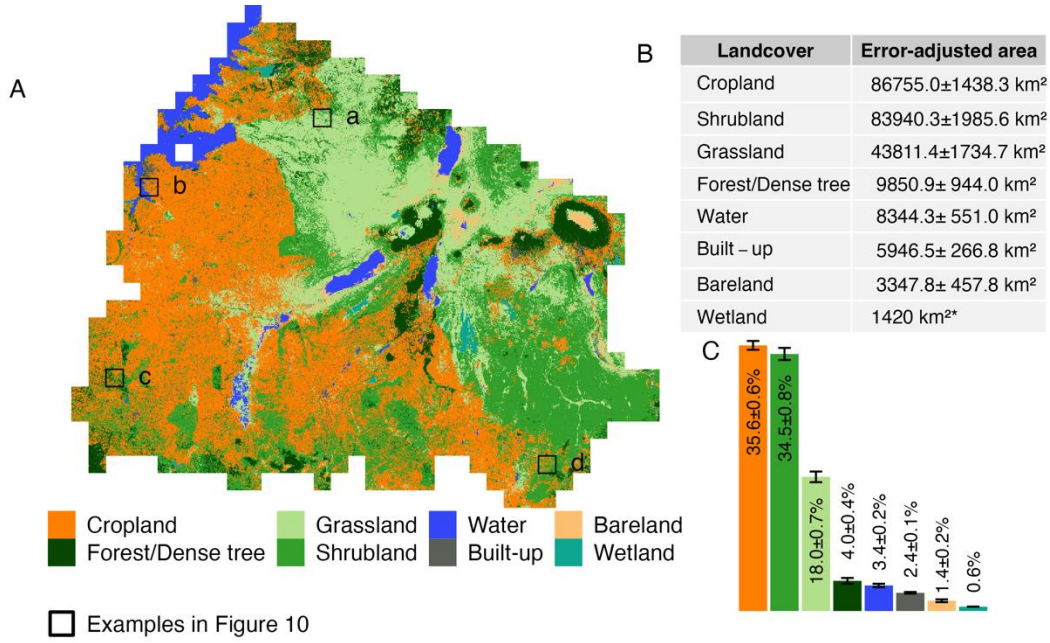

Figure 1-9. Predicted land cover (A), error-adjusted area estimates (B), and proportion of each land cover type (C) with 95% confidence interval over the study area (*The estimated area of wetland is directly calculated by the rasterized OSM layer without error adjustment)

In northern Tanzania, the predominant land cover types are cropland, shrubland, and grassland (Figure 1-9B & 9C). The error-adjusted area estimate (Olofsson et al., 2014) with a 95% confidence interval of the cropland is 86755±1438.3 km^2; shrubland is 83940.3±1985.6 km^2 and grassland is 43811.4±1734.7 km^2. Most of the cropland is distributed in the Central plateau agro-ecological zone, with plains and arable lands (Figure 1-1 & Figure 1-9A). The southern part of this zone is less cultivated and is largely covered by shrubs. Shrubland are found primarily in the Eastern plateau zone with a moist climate, where the land is mainly uncultivated because it is

falls within protected areas. Grasslands predominate in the Northern riftzone and volcanic highlands, where the Serengeti National Park, Ngorongoro Conservation Area, and Arusha National Park are located.

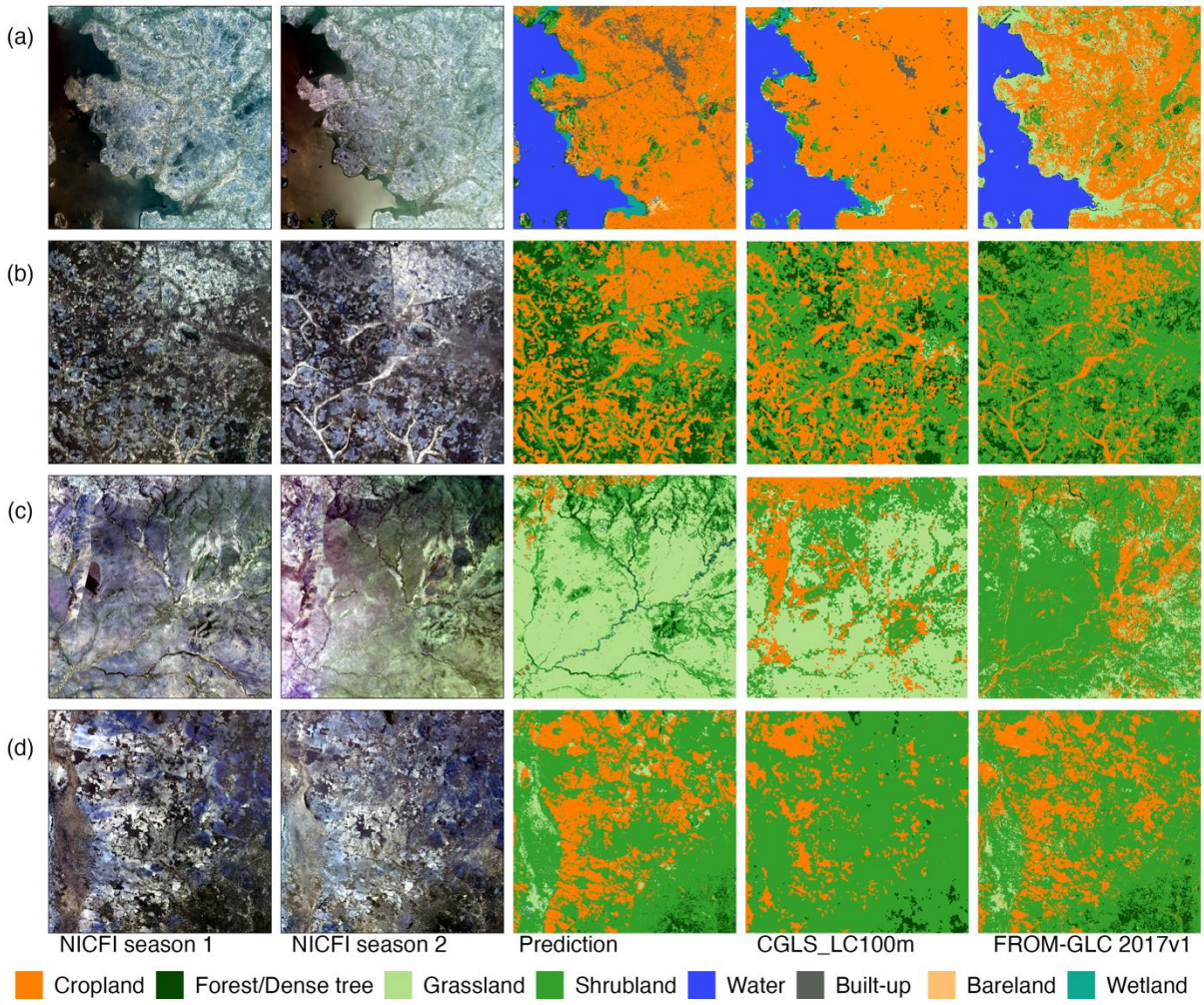

Figure 1-10. Seasonal NICFI images, prediction in this study, and the CGLS_LC100m and FROM-GLC 2017v1 product of four example tiles (a and b are in Central plateau agro-ecological zone, c is in Northern Riftzone and Volcanic highlands zone, and d is in Eastern plateau zone)

Comparisons of our land cover maps within different agro-ecological zones with the two existing multi-class land cover products used to make consensus labels show that our predictions inherited the advantages and mitigated the errors contained within the original land

cover products. For example, Figure 1-10a shows that our model reduced the substantial omission error in built-up areas and shrubland in the other two products while maintaining the broad distribution of croplands, and Figure 1-10b shows that our updated map more effectively captures forests and croplands. Figure 1-10c shows that CGLS_LC overestimated cropland while FROM-GLC overestimated shrubs, and our prediction reduced these issues significantly. Similarly, in Figure 1-10d, CGLS_LC underestimated cropland and our prediction detected cropland as effectively as FROM-GLC. Owing to the high resolution of satellite imagery, our predictions reveal more landscape details and delineate more precise object boundaries. For instance, the dense tree cover within riverine forests is clearly delineated (Figure 1-10c), as well as scattered residential areas (Figure 1-10a).

1.5 Discussion

Our results demonstrate an effective and operationalizable approach for mapping land cover in data sparse areas, which is capable of producing greatly improved land cover data in hard to map tropical savannas, where up-to-date and accurate land change information is critical given the rapid pace of change (Bullock et al., 2021) and the relative inaccuracy and infrequent production of existing land cover products. The resulting map has noticeable improvements in land cover type differentiation relative to existing land cover products, particularly in cropland and grassland. These gains are in part attributable to the integration of high-resolution PlanetScope NICFI basemaps, which significantly enhances object boundary delineation, and improves detection of minor land cover types (e.g., residential) and the description of landscape patches. The Sentinel-1 time series provides a cloud-free temporal signature of landscape

features, which can overcome the spectral limitations of optical satellite imagery (e.g., PlanetScope) and increase the ability to distinguish different land cover types (Jacob et al., 2020).

A key feature of our approach that enables its operationalization is its ability to generate a large number of labels without intensive human effort, which it does by leveraging existing land cover datasets. Even though these data individually have substantial problems and often perform poorly in our study area (Table A-2), particularly because they are developed over regional to global extents (Buchhorn et al., 2020b; Congalton et al., 2017; Gong et al., 2019; Walsh et al., 2018), using the consensus of these products can provide a useful starting point for collecting a large number of training samples that can be synthetically improved through an easily implemented machine learning approach. As a widely used algorithm, Random Forest performs well for land cover classification (Sheykhmousa et al., 2020; Talukdar et al., 2020), has interpretable structure, and provides additional variable importance information that can be used to inform further modeling efforts. It can be trained using both point- and image-based training samples, which makes the model easy to rapidly develop and apply. However, the algorithm cannot learn from the spatial patterns of features or their relationships (Breiman, 2001; Chan and Paelinckx, 2008), thus it is necessary to carefully select additional supplementary features that can improve model performance. Neural network architectures, such as U-Net, which can learn from contextual patterns in imagery to improve object boundary delineation (Iglovikov et al., 2017; Rakhlin et al., 2018), and thereby overcome one of Random Forest's major limitations. However, this improved performance comes at the expense of collecting larger volumes of training data, consisting of chips that are fully labelled to provide the model with the necessary spatial context. Even though strategies such as weak supervision (Wang et al., 2020) have been

proposed to reduce labelling effort by using fragmented labels, the lack of precise localization information can lead to the failure of small size objects, particularly across complex fragmented landscapes. Using two models together as we have done here plays to each model's strengths while helping to overcome the major challenge of limited training data.

There, nevertheless, is still room for improvement in both approach and datasets. In our study, we simply resampled S1 SAR imagery to align with PlanetScope NICFI basemaps in order to make full use of the high resolution and reduce the risk of adding more uncertainties due to extra processing. In fact, multiple methods better than resampling exist to deal with imagery with different scales. The time intervals for semi-annual NICFI basemaps are not customized regionally, which may not match with local seasonality. In this study, we selected NICFI scenes (2017-12-01 – 2018-05-30 and 2018-06-01 – 2018-11-30) covering dates that generally align with the agriculture year (2017-10-01 – 2018-09-30). Because PlanetScope basemaps were mosaiced every six months, we assumed a two-month shift would not cause any significant issues in dynamic landscapes. This deficiency, however, can limit the mapping ability that relies on temporal signatures in spectra (e.g., mapping crop types).

In the Sub-Saharan African landscape, fallow is a common and critical land-use type, but it is usually not a target class in coarse level land cover studies. The fallow is often interpreted as farmland due to its regular shape and close location to active farmland. However, it has more similar spectral and temporal features to grassland or bareland (Tong et al., 2020). This could be a significant reason for misclassification.

1.6 Conclusion

We provided an operational workflow for rapid land cover mapping that can be applied to heterogeneous landscapes without good ground truth references. The proposed workflow ensembles not only multiple LC products but also artificial and human intelligence to increase classification reliability. We then demonstrated its capacity to map a complex tropical savanna landscape in Northern Tanzania. The resultant land cover map can be used in investigating agricultural expansion and development, analyzing deforestation or sustainability of protected areas, and many other ecological applications.

Additionally, we provided a high-resolution land cover dataset with label quality information at large scales across the savanna landscape. It well presents features, structures, and distribution of major land cover categories in our study area. Therefore, they can be used to train different types of land cover models, and more importantly to investigate how to improve land cover mapping with noisy labels that are inevitable in the real world. The pre-trained U-Net network upon this dataset can directly be used to map land cover if the landscape condition is similar and input datasets are available. It can also be fine-tuned and transferred to other cases or years.

The land-cover dataset and land cover map produced in our study are accessible under https://osf.io/4qj36/. Code of workflow and of pre-processing satellite image for this article can be found on GitHub at https://github.com/LLeiSong/hrlcm and https://github.com/LLeiSong/sentinelPot.

CRediT authorship contribution statement

Lei Song: Conceptualization, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Funding acquisition. **Anna B. Estes**: Validation, Writing - Review & Editing. **Lyndon D. Estes**: Conceptualization, Methodology, Resources, Supervision, Writing - Review & Editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Future Investigators in NASA Earth and Space Science and Technology (FINESST) program (award number: 80NSSC20K1640).

References

Abdi, A.M., Brandt, M., Abel, C., Fensholt, R., 2022. Satellite Remote Sensing of Savannas: Current Status and Emerging Opportunities. J. Remote Sens. 2022, 1–20. https://doi.org/10.34133/2022/9835284

Anderson, S.J., Ankor, B.L., Sutton, P.C., 2017. Ecosystem service valuations of South Africa using a variety of land cover data sources and resolutions. Ecosyst. Serv. 27, 173–178. https://doi.org/10.1016/j.ecoser.2017.06.001

Aquino, C., Mitchard, E.T.A., McNicol, I.M., Carstairs, H., Burt, A., Puma Vilca, B.L., Mayta, S., Disney, M., 2022. Detecting Tropical Forest Degradation Using Optical Satellite Data: An Experiment in Peru Show Texture at 3 M Gives Best Results. https://doi.org/10.20944/preprints202202.0141.v1

Awuah, K.T., Aplin, P., 2021. Fusion of Sentinel-2 Data with High Resolution Open Access Planet Basemaps for Grazing Lawn Detection in Southern African Savannahs, in: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. IEEE, pp. 1409–1412.

Barsi, Á., Kugler, Zs., László, I., Szabó, Gy., Abdulmutalib, H.M., 2018. Accuracy Dimensions in Remote Sensing. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. XLII–3, 61–67. https://doi.org/10.5194/isprs-archives-XLII-3-61-2018

Bengio, Y., 2012. Practical recommendations for gradient-based training of deep architectures, in: Neural Networks: Tricks of the Trade. Springer, pp. 437–478.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The balanced accuracy and its posterior distribution, in: 2010 20th International Conference on Pattern Recognition. IEEE, pp. 3121–3124.

Buchhorn, M., Lesiv, M., Tsendbazar, N.-E., Herold, M., Bertels, L., Smets, B., 2020a. Copernicus global land cover layers—collection 2. Remote Sens. 12, 1044.

Buchhorn, M., Smets, B., Bertels, L., De Roo, B., Lesiv, M., Tsendbazar, N.-E., Herold, M., Fritz, S., 2020b. Copernicus Global Land Service: Land Cover 100m: Collection 3 Epoch 2018, Globe. Version V3 01Data Set.

Bullock, E.L., Healey, S.P., Yang, Z., Oduor, P., Gorelick, N., Omondi, S., Ouko, E., Cohen, W.B., 2021. Three Decades of Land Cover Change in East Africa. Land 10, 150. https://doi.org/10.3390/land10020150

Burke, M., Lobell, D.B., 2017. Satellite-based assessment of yield variation and its determinants in smallholder African systems. Proc. Natl. Acad. Sci. 114, 2189–2194. https://doi.org/10.1073/pnas.1616919114

Campos-Taberner, M., García-Haro, F.J., Martínez, B., Izquierdo-Verdiguier, E., Atzberger, C., Camps-Valls, G., Gilabert, M.A., 2020. Understanding deep learning in land use

classification based on Sentinel-2 time series. Sci. Rep. 10, 17188. https://doi.org/10.1038/s41598-020-74215-5

Chamorro Martinez, J.A., Cué La Rosa, L.E., Feitosa, R.Q., Sanches, I.D., Happ, P.N., 2021. Fully convolutional recurrent networks for multidate crop recognition from multitemporal image sequences. ISPRS J. Photogramm. Remote Sens. 171, 188–201. https://doi.org/10.1016/j.isprsjprs.2020.11.007

Chan, J.C.-W., Paelinckx, D., 2008. Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. Remote Sens. Environ. 112, 2999–3011. https://doi.org/10.1016/j.rse.2008.02.011

Cheng, Y., Vrieling, A., Fava, F., Meroni, M., Marshall, M., Gachoki, S., 2020. Phenology of short vegetation cycles in a Kenyan rangeland from PlanetScope and Sentinel-2. Remote Sens. Environ. 248, 112004. https://doi.org/10.1016/j.rse.2020.112004

Congalton, R., Yadav, K., McDonnell, K., Poehnelt, J., Stevens, B., Gumma, M., Teluguntla, P., Thenkabail, P., 2017. Global Food Security-support Analysis Data (GFSAD) Cropland Extent 2015 Validation 30 m V001.

Doggart, N., Morgan-Brown, T., Lyimo, E., Mbilinyi, B., Meshack, C.K., Sallu, S.M., Spracklen, D.V., 2020. Agriculture is the main driver of deforestation in Tanzania. Environ. Res. Lett. 15, 034028. https://doi.org/10.1088/1748-9326/ab6b35

Elmes, A., Alemohammad, H., Avery, R., Caylor, K., Eastman, J., Fishgold, L., Friedl, M., Jain, M., Kohli, D., Laso Bayas, J., Lunga, D., McCarty, J., Pontius, R., Reinmann, A., Rogan, J., Song, L., Stoynova, H., Ye, S., Yi, Z.-F., Estes, L., 2020. Accounting for Training Data Error in Machine Learning Applied to Earth Observations. Remote Sens. 12, 1034. https://doi.org/10.3390/rs12061034

Estes, L.D., Ye, S., Song, L., Luo, B., Eastman, J.R., Meng, Z., Zhang, Q., McRitchie, D., Debats, S.R., Muhando, J., Amukoa, A.H., Kaloo, B.W., Makuru, J., Mbatia, B.K., Muasa, I.M., Mucha, J., Mugami, A.M., Mugami, J.M., Muinde, F.W., Mwawaza, F.M., Ochieng, J., Oduol, C.J., Oduor, P., Wanjiku, T., Wanyoike, J.G., Avery, R.B., Caylor, K.K., 2022. High Resolution, Annual Maps of Field Boundaries for Smallholder-Dominated Croplands at National Scales. Front. Artif. Intell. 4, 744863. https://doi.org/10.3389/frai.2021.744863

Fritz, S., See, L., Rembold, F., 2010. Comparison of global and regional land cover maps with statistical information for the agricultural domain in Africa. Int. J. Remote Sens. 31, 2237–2256. https://doi.org/10.1080/01431160902946598

Fritz, S., You, L., Bun, A., See, L., McCallum, I., Schill, C., Perger, C., Liu, J., Hansen, M., Obersteiner, M., 2011. Cropland for sub-Saharan Africa: A synergistic approach using five land cover data sets. Geophys. Res. Lett. 38.

Gómez, C., White, J.C., Wulder, M.A., 2016. Optical remotely sensed time series data for land cover classification: A review. ISPRS J. Photogramm. Remote Sens. 116, 55–72. https://doi.org/10.1016/j.isprsjprs.2016.03.008

Gong, P., Liu, H., Zhang, M., Li, C., Wang, J., Huang, H., Clinton, N., Ji, L., Li, Wenyu, Bai, Y., Chen, B., Xu, B., Zhu, Z., Yuan, C., Ping Suen, H., Guo, J., Xu, N., Li, Weijia, Zhao, Y., Yang, J., Yu, C., Wang, X., Fu, H., Yu, L., Dronova, I., Hui, F., Cheng, X., Shi, X., Xiao, F., Liu, Q., Song, L., 2019. Stable classification with limited sample: transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017. Sci. Bull. 64, 370–373. https://doi.org/10.1016/j.scib.2019.03.002

Houborg, R., McCabe, M.F., 2018. A Cubesat enabled Spatio-Temporal Enhancement Method (CESTEM) utilizing Planet, Landsat and MODIS data. Remote Sens. Environ. 209, 211–226. https://doi.org/10.1016/j.rse.2018.02.067

Huete, A.R., 1988. A soil-adjusted vegetation index (SAVI). Remote Sens. Environ. 25, 295–309. https://doi.org/10.1016/0034-4257(88)90106-X

Iglovikov, V., Mushinskiy, S., Osin, V., 2017. Satellite Imagery Feature Detection using Deep Convolutional Neural Network: A Kaggle Competition.

Jacob, A.W., Vicente-Guijalba, F., Lopez-Martinez, C., Lopez-Sanchez, J.M., Litzinger, M., Kristen, H., Mestre-Quereda, A., Ziolkowski, D., Lavalle, M., Notarnicola, C., Suresh, G., Antropov, O., Ge, S., Praks, J., Ban, Y., Pottier, E., Mallorqui Franquet, J.J., Duro, J., Engdahl, M.E., 2020. Sentinel-1 InSAR Coherence for Land Cover Mapping: A Comparison of Multiple Feature-Based Classifiers. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 535–552. https://doi.org/10.1109/JSTARS.2019.2958847

Jiang, Z., Huete, A.R., Didan, K., Miura, T., 2008. Development of a two-band enhanced vegetation index without a blue band. Remote Sens. Environ. 112, 3833–3845.

Jin, Z., Azzari, G., You, C., Di Tommaso, S., Aston, S., Burke, M., Lobell, D.B., 2019. Smallholder maize area and yield mapping at national scales with Google Earth Engine. Remote Sens. Environ. 228, 115–128. https://doi.org/10.1016/j.rse.2019.04.016

Kaufman, Y.J., Tanre, D., 1992. Atmospherically resistant vegetation index (ARVI) for EOS-MODIS. IEEE Trans. Geosci. Remote Sens. 30, 261–270.

Kerner, H., Tseng, G., Becker-Reshef, I., Nakalembe, C., Barker, B., Munshell, B., Paliyam, M., Hosseini, M., 2020. Rapid Response Crop Maps in Data Sparse Regions. ArXiv200616866 Cs Eess.

Laso Bayas, J.C., Lesiv, M., Waldner, F., Schucknecht, A., Duerauer, M., See, L., Fritz, S., Fraisl, D., Moorthy, I., McCallum, I., Perger, C., Danylo, O., Defourny, P., Gallego, J., Gilliams, S., Akhtar, I. ul H., Baishya, S.J., Baruah, M., Bungnamei, K., Campos, A., Changkakati, T., Cipriani, A., Das, Krishna, Das, Keemee, Das, I., Davis, K.F., Hazarika, P., Johnson, B.A., Malek, Z., Molinari, M.E., Panging, K., Pawe, C.K., Pérez-Hoyos, A., Sahariah, P.K., Sahariah, D., Saikia, A., Saikia, M., Schlesinger, P., Seidacaru, E., Singha, K., Wilson, J.W., 2017. A global reference database of crowdsourced cropland

data collected using the Geo-Wiki platform. Sci. Data 4, 170136. https://doi.org/10.1038/sdata.2017.136

Leite-Filho, A.T., Soares-Filho, B.S., Davis, J.L., Abrahão, G.M., Börner, J., 2021. Deforestation reduces rainfall and agricultural revenues in the Brazilian Amazon. Nat. Commun. 12, 2591. https://doi.org/10.1038/s41467-021-22840-7

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440.

Luo, L., Xiong, Y., Liu, Y., Sun, X., 2019. Adaptive Gradient Methods with Dynamic Bound of Learning Rate. ArXiv190209843 Cs Stat.

Moody, A., Johnson, D.M., 2001. Land-Surface Phenologies from AVHRR Using the Discrete Fourier Transform. Remote Sens. Environ. 75, 305–323. https://doi.org/10.1016/S0034-4257(00)00175-9

Norway's International Climate and Forest Initiative (NICFI) [WWW Document], 2020. . NICFI. URL https://www.nicfi.no/ (accessed 4.17.22).

Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E., Wulder, M.A., 2014. Good practices for estimating area and assessing accuracy of land change. Remote Sens. Environ. 148, 42–57. https://doi.org/10.1016/j.rse.2014.02.015

Pascual, A., Tupinambá-Simões, F., de Conto, T., 2022. Using multi-temporal tree inventory data in eucalypt forestry to benchmark global high-resolution canopy height models. A showcase in Mato Grosso, Brazil. Ecol. Inform. 70, 101748. https://doi.org/10.1016/j.ecoinf.2022.101748

Perez, L., Wang, J., 2017. The effectiveness of data augmentation in image classification using deep learning. ArXiv Prepr. ArXiv171204621.

Pérez-Hoyos, A., Udías, A., Rembold, F., 2020. Integrating multiple land cover maps through a multi-criteria analysis to improve agricultural monitoring in Africa. Int. J. Appl. Earth Obs. Geoinformation 88, 102064. https://doi.org/10.1016/j.jag.2020.102064

Pettorelli, N., Wegmann, M., Skidmore, A., Mücher, S., Dawson, T.P., Fernandez, M., Lucas, R., Schaepman, M.E., Wang, T., O'Connor, B., Jongman, R.H.G., Kempeneers, P., Sonnenschein, R., Leidner, A.K., Böhm, M., He, K.S., Nagendra, H., Dubois, G., Fatoyinbo, T., Hansen, M.C., Paganini, M., de Klerk, H.M., Asner, G.P., Kerr, J.T., Estes, A.B., Schmeller, D.S., Heiden, U., Rocchini, D., Pereira, H.M., Turak, E., Fernandez, N., Lausch, A., Cho, M.A., Alcaraz-Segura, D., McGeoch, M.A., Turner, W., Mueller, A., St-Louis, V., Penner, J., Vihervaara, P., Belward, A., Reyers, B., Geller, G.N., 2016. Framing the concept of satellite remote sensing essential biodiversity variables: challenges and future directions. Remote Sens. Ecol. Conserv. 2, 122–131. https://doi.org/10.1002/rse2.15

Planet Team, 2017. Planet application program interface: In space for life on Earth. San Franc. CA 2017, 40.

Rakhlin, A., Davydow, A., Nikolenko, S., 2018. Land Cover Classification from Satellite Imagery with U-Net and Lovász-Softmax Loss, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, Salt Lake City, UT, USA, pp. 257–2574. https://doi.org/10.1109/CVPRW.2018.00048

Reiner, F., Brandt, M., Tong, X., Skole, D., Kariryaa, A., Ciais, P., Davies, A., Hiernaux, P., Chave, J., Mugabowindekwe, M., Igel, C., Oehmcke, S., Gieseke, F., Li, S., Liu, S., Saatchi, S., Boucher, P., Singh, J., Taugourdeau, S., Dendoncker, M., Song, X.-P., Mertz, O., Tucker, C.J., Fensholt, R., 2022. More than one quarter of Africa's tree cover found outside areas previously classified as forest. https://doi.org/10.21203/rs.3.rs-1816495/v2

Ren, H., Liu, Y., Chang, X., Yang, J., Xiao, X., Huang, X., 2022. Mapping High-Resolution Global Impervious Surface Area: Status and Trends. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 1–21. https://doi.org/10.1109/JSTARS.2022.3201380

Rienow, A., Schweighöfer, J., Dedring, T., Goebel, M., Graw, V., 2022. Detecting land use and land cover change on Barbuda before and after the Hurricane Irma with respect to potential land grabbing: A combined volunteered geographic information and multi sensor approach. Int. J. Appl. Earth Obs. Geoinformation 108, 102732. https://doi.org/10.1016/j.jag.2022.102732

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Roy, D.P., Huang, H., Houborg, R., Martins, V.S., 2021. A global analysis of the temporal availability of PlanetScope high spatial resolution multi-spectral imagery. Remote Sens. Environ. 264, 112586. https://doi.org/10.1016/j.rse.2021.112586

Rufin, P., Bey, A., Picoli, M., Meyfroidt, P., 2022. Large-area mapping of active cropland and short-term fallows in smallholder landscapes using PlanetScope data. Int. J. Appl. Earth Obs. Geoinformation 112, 102937. https://doi.org/10.1016/j.jag.2022.102937

Schmitt, M., Hughes, L.H., Qiu, C., Zhu, X.X., 2019. SEN12MS – A CURATED DATASET OF GEOREFERENCED MULTI-SPECTRAL SENTINEL-1/2 IMAGERY FOR DEEP LEARNING AND DATA FUSION. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. IV-2/W7, 153–160. https://doi.org/10.5194/isprs-annals-IV-2-W7-153-2019

Sebastian, K., 2009. Agro-ecological Zones of Africa.

Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P., Homayouni, S., 2020. Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 6308–6325. https://doi.org/10.1109/JSTARS.2020.3026724

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. J. Big Data 6, 1–48.

Smith, L.N., 2017. Cyclical learning rates for training neural networks, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 464–472.

Solbrig, O.T., 1996. The diversity of the savanna ecosystem, in: Biodiversity and Savanna Ecosystem Processes. Springer, pp. 1–27.

Solórzano, J.V., Mas, J.F., Gao, Y., Gallardo-Cruz, J.A., 2021. Land Use Land Cover Classification with U-Net: Advantages of Combining Sentinel-1 and Sentinel-2 Imagery. Remote Sens. 13, 3600. https://doi.org/10.3390/rs13183600

Song, X.-P., Hansen, M.C., Stehman, S.V., Potapov, P.V., Tyukavina, A., Vermote, E.F., Townshend, J.R., 2018. Global land change from 1982 to 2016. Nature 560, 639–643. https://doi.org/10.1038/s41586-018-0411-9

Sugimoto, R., Kato, S., Nakamura, R., Tsutsumi, C., Yamaguchi, Y., 2022. Deforestation detection using scattering power decomposition and optimal averaging of volume scattering power in tropical rainforest regions. Remote Sens. Environ. 275, 113018. https://doi.org/10.1016/j.rse.2022.113018

Sun, R.-Y., 2020. Optimization for deep learning: An overview. J. Oper. Res. Soc. China 8, 249–294.

Talukdar, S., Singha, P., Mahato, S., Shahfahad, Pal, S., Liou, Y.-A., Rahman, A., 2020. Land-Use Land-Cover Classification by Machine Learning Classifiers for Satellite Observations—A Review. Remote Sens. 12, 1135. https://doi.org/10.3390/rs12071135

Tong, X.-Y., Xia, G.-S., Lu, Q., Shen, H., Li, S., You, S., Zhang, L., 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. Remote Sens. Environ. 237, 111322. https://doi.org/10.1016/j.rse.2019.111322

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., others, 2012. GMES Sentinel-1 mission. Remote Sens. Environ. 120, 9–24.

Torrey, L., Shavlik, J., 2010. Transfer learning, in: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques. IGI global, pp. 242–264.

Tsalyuk, M., Kelly, M., Getz, W.M., 2017. Improving the prediction of African savanna vegetation variables using time series of MODIS products. ISPRS J. Photogramm. Remote Sens. 131, 77–91. https://doi.org/10.1016/j.isprsjprs.2017.07.012

Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens. Environ. 8, 127–150.

Vizzari, M., 2022. PlanetScope, Sentinel-2, and Sentinel-1 Data Integration for Object-Based Land Cover Classification in Google Earth Engine. Remote Sens. 14, 2628. https://doi.org/10.3390/rs14112628

Volpi, M., Tuia, D., 2016. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. IEEE Trans. Geosci. Remote Sens. 55, 881–893.

Walsh, M., Meliyo, J., Awiti, A., Scott, B., Walsh, B., Macmillan, B., 2018. Tanzania Soil Information Service (TanSIS).

Wang, S., Chen, W., Xie, S.M., Azzari, G., Lobell, D.B., 2020. Weakly Supervised Deep Learning for Segmentation of Remote Sensing Imagery. Remote Sens. 12, 207. https://doi.org/10.3390/rs12020207

Whitley, R., Beringer, J., Hutley, L.B., Abramowitz, G., De Kauwe, M.G., Evans, B., Haverd, V., Li, L., Moore, C., Ryu, Y., others, 2017. Challenges and opportunities in land surface modelling of savanna ecosystems. Biogeosciences 14, 4711–4732.

Xu, Y., Yu, L., Feng, D., Peng, D., Li, C., Huang, X., Lu, H., Gong, P., 2019. Comparisons of three recent moderate resolution African land cover datasets: CGLS-LC100, ESA-S2-LC20, and FROM-GLC-Africa30. Int. J. Remote Sens. 40, 6185–6202.

Yang, J., Huang, X., 2021. 30 m annual land cover and its dynamics in China from 1990 to 2019. Earth Syst Sci Data Discuss 2021, 1–29.

Zhang, W., Brandt, M., Wang, Q., Prishchepov, A.V., Tucker, C.J., Li, Y., Lyu, H., Fensholt, R., 2019. From woody cover to woody canopies: How Sentinel-1 and Sentinel-2 data advance the mapping of woody plants in savannas. Remote Sens. Environ. 234, 111465. https://doi.org/10.1016/j.rse.2019.111465

Zhang, X., Liu, L., Chen, X., Gao, Y., Xie, S., Mi, J., 2021. GLC_FCS30: global land-cover product with fine classification system at 30 m using time-series Landsat imagery. Earth Syst. Sci. Data 13, 2753–2776. https://doi.org/10.5194/essd-13-2753-2021

Chapter 2

itsdm: Isolation Forest-based presence-only species distribution modeling and explanation in R

Lei Song [a*], Lyndon Despard Estes [a]

[a] Graduate School of Geography, Clark University, Worcester, MA, USA

[*]Corresponding author: Lei Song

Published as:

Song, L., & Estes, L. (2023). *itsdm*: Isolation forest-based presence-only species distribution modelling and explanation in R. *Methods in Ecology and Evolution*.

2.1 Abstract

1. Multiple statistical algorithms have been used for species distribution modeling (SDM). Due to shortcomings in species occurrence datasets, presence-only methods (such as MaxEnt) have become increasingly widely used. However, sampling bias remains a challenging issue, particularly for density-based approaches. The Isolation Forest (iForest) algorithm is a presence-only method less sensitive to sampling patterns and over-fitting because it fits the model by describing the unsuitable instead of suitable conditions.
2. Here we present the *itsdm* package for species distribution modeling with iForest, which provides a workflow wrapper for the algorithms in iForest family and convenient tools for model diagnostic and post-modeling analysis.

3. *itsdm* allows users to fit and evaluate an iForest SDM using presence-only occurrence data. It also helps the users to understand relationships between species and the living environment using Shapley values, a suggested technique in explainable artificial intelligence (xAI). Additionally, *itsdm* can make spatial response maps that indicate how species respond to environmental variables across space and detect areas potentially affected by a changing environment.
4. We demonstrated the usage of the *itsdm* package and compared iForest with other mainstream SDMs using virtual species. The results enlightened that iForest is an advantageous presence-only SDM when the actual distribution range is unclear.

KEYWORDS: Presence-only, Species distribution modeling (SDM), Isolation Forest, Shapley values, Explainable artificial intelligence (xAI)

2.2 Introduction

Statistical methods and associated algorithms have been used for decades to develop species distribution models (SDMs) (Guisan & Zimmermann, 2000) because of their practical usefulness in ecological decision-making and conservation planning. Their usage continues to expand as ever-growing accessibility of occurrence data from public databases (Sofaer et al., 2019). However, many species occurrence datasets were not gathered in structured surveys. They thus may partially cover suitable habitats, contain sampling issues, and often lack absence cases, which negatively impact SDMs (Beck et al., 2014). Background sampling is a common way to deal with the missing absence records, but without enough background knowledge and proper sampling strategies, taking background samples as pseudo-absence records may confound the

environmental response of the modeled species (Barbet-Massin et al., 2012). Anomaly detection algorithms, which are semi-supervised, can take presence-only records without any pseudo-absence samples, reducing the risk of adding false information. Among them, maximum entropy (MaxEnt) has risen to dominance and has been widely used in many case studies (Elith et al., 2011; Phillips & Dudík, 2008). However, as with other density-based methods, it is easily affected by sampling patterns and overfits (Kramer-Schadt et al., 2013; Merow et al., 2013; Radosavljevic & Anderson, 2014). The regularization multiplier may reduce the issues by controlling model complexity, but the effects are species-specific and prone to sample size (Morales et al., 2017; Radosavljevic & Anderson, 2014). Isolation Forest (iForest) uses a novel approach based on the depth of the branch in the tree to calculate the probabilities (Liu et al., 2008, 2010, 2012). Optimizing to unsuitable conditions and not relying on sample density to fit the model, it thus suffers less from overfitting and sampling issues. The tree structure functions similarly to profile models that describe the species-environment relationship as an "environmental profile" (Franklin, 2010), and consequently trends to predict the environmental suitability rather than the probability of detection.

In computer science, iForest is widely applied in spatial and non-spatial anomaly detection and one-class classification problems (Feremans et al., 2020; Khan et al., 2019; Li et al., 2019), but it has not yet been adopted widely in ecology relatively because iForest lacks a standard toolkit to assess ecological validity and produce detailed summaries of fitted relationships. Many post-hoc methods have been proposed to analyze the behaviors of models with non-interpretable structures. The "evaluation strip" technique (Elith et al., 2005) can visualize variable responses for any modeling approach, aiding the users in evaluating and comparing models with different structures. Phillips et al. (2006) implemented leave out one

Jackknife test to identify variables with significant individual effects. Shapley values technique is listed as one of the post-hoc model-agnostic tools in explainable artificial intelligence (xAI) and is encouraged to be applied in SDM research domain (Ryo et al., 2021). It can explain the relative contribution of each feature to the prediction at a given instance locally and summarize variable response and variable importance globally (Lundberg & Lee, 2017; Shapley, 1953).

To take advantage of the ability of iForest to handle presence-only data with less sensitivity to sampling patterns, we developed a new R (R Core Team, 2021) package *itsdm*. It provided a wrapper for iForest and its related variants (Cortes, 2021b, 2022; Guha et al., 2016; Hariri et al., 2019; Liu et al., 2008, 2010) to do species distribution modeling, alongside methods delivering ecological insights from the model. This package aims to provide ecological modelers with an additional tool for creating SDMs, which can complement well-established existing approaches, such as those implemented in BIOMOD (Thuiller et al., 2009, 2021).

2.3 Package structure and description

itsdm is a workflow wrapper coded in R and knits iForest and Shapley value explanation into an SDM workflow. The package's functions are in four groups (Table 2-1): pre-modeling analysis, modeling, model explanation, and post-modeling analysis. The pre-modeling analysis functions diagnose the relationships between environmental variables and target potential sampling errors in the occurrence dataset. The model implementation functions format observation dataset, build and evaluate the model with different user settings. The model explanation functions delineate the importance of environmental variables and the species' spatial and non-spatial responses to them. The package also contains a post-modeling toolkit for

further analysis of modeling results, for instance, analyzing the impacts of a changing environmental variable, converting predicted suitability to a presence-absence, and comparing the contribution of environmental variables to observations. Importantly, all Shapley values-based functions (such as *shap_dependence, shap_spatial_response, detect_envi_change, and variable_contrib*) can apply to any fitted models as long as the function inputs are correctly set (see example in section 2.5.2). Because visualization is critical in ecological modeling, *itsdm* provides corresponding *print* and/or *plot* generic functions to visualize every object (Table 2-1).

Table 2-1 Core functions and descriptions in *itsdm*

Function (Object)		Visualization	Description
Pre-modeling analysis	dim_reduce (ReducedImageStack)	print	Select numeric environmental variables with pairwise Pearson correlation lower than a defined threshold. The user can specify preferred variables.
	suspicious_env_outliers (EnvironmentalOutlier)	print, plot	Detect suspicious environmental outliers in the occurrence dataset according to each environmental covariate's general condition.
Modeling	format_observation (FormatOccurrence)	print	Quickly format the dataset to fit into the *itsdm* workflow.
	isotree_po (POIsotree)	print	Build Isolation Forest-based SDM and do the related model explanation, which optionally calls model explanation functions.
	evaluate_po (POEvaluation)	print, plot	Evaluate the model based on presence-only data.
Model explanation	variable_analysis (VariableAnalysis)	print, plot	Evaluate environmental variable importance using leave one out Jackknife test and Shapley values.
	marginal_response (MarginalResponse)	plot	Calculate marginal response curve of environmental variables using the evaluation strip method proposed by Elith et al. (2005).
	independent_response (IndependentResponse)	plot	Calculate independent response curve of environmental variables by creating an independent model using only one variable each time (Phillips et al., 2006).
	shap_dependence (ShapDependence)	plot	Calculate Shapley value-based variable dependence plot which is introduced in section 2.5. It is a supplementary plot for marginal and independent response curves using a completely different method. It also can explore the relationship between two environmental variables.
	spatial_response (SpatialResponse)	plot	Generate spatially partial dependence maps with type of marginal, independent, and Shapley value based.
	shap_spatial_response (SHAPSpatial)	plot	Generate spatially partial dependence maps only with Shapley values. It can work on other external models.

Post-modeling analysis	detect_envi_change (EnviChange)	print, plot	Use post-hoc Shapley values technique to detect the tipping points and potentially affected areas due to a changing environmental variable.
	convert_to_pa (PAConversion)	print, plot	Convert predicted suitability to presence-absence map.
	variable_contrib (VariableContribution)	print, plot	Evaluate local and/or global variable contributions for interested observation(s).

2.4 SDMs with isolation forest

Isolation Forest (iForest) is built based on the decision tree architecture to distinguish anomalies or outliers from a set of samples (e.g. presence-only samples). Because the majority of the samples are normal, anomalies are few and different. In presence-only SDM, it means samples gathered in less suitable areas are lower in quantity and environmentally different from samples in suitable areas. iForest aims to fit a model to describe these anomalies rather than the normal samples, therefore, does not necessarily need background samples. More importantly, it is more robust to sampling issues and overfitting.

iForest uses the path in the tree structure to calculate the probability of a sample being anomalous (termed as anomaly score). Reflecting on a tree structure, the anomalies are isolated closer to the tree's root note, so they have shorter paths (Figure 2-1a). Given an isolation tree built on a dataset $X = \{x_1, \ldots, x_n\}$ of n samples, X is divided recursively by a test for every internal-node T_{in} to two sibling nodes (T_l, T_r) until the node becomes an external-node T_{ex} with no child or a predefined depth limit is reached. Within the feature space, the test is a hyperplane defined by a random normal vector and intercept (Figure 2-1b). The node splitting criterion for a given point $\vec{x}$ is as follows (Hariri et al., 2019):

$$(\vec{x} - \vec{p}) \cdot \vec{n} \leq 0 \tag{2-1}$$

Where $\vec{n}$ is a random normal vector uniformly over the unit N-Sphere which specifies the $N-1$ dimensional hyperplane to split nodes for a dataset with N attributes (N-dimensional). $\vec{p}$ is a set of values from a uniform distribution over the range of possible values at each node which serves as a set of random intercepts of the hyperplane. This is a general definition of iForest called Extended Isolation Forest (EIF). Standard iForest is a special case of EIF whose split test only consists of a randomly-selected attribute q and a split value p such that the test $q < p$ splits the node into sibling nodes (Liu et al., 2008, 2012).

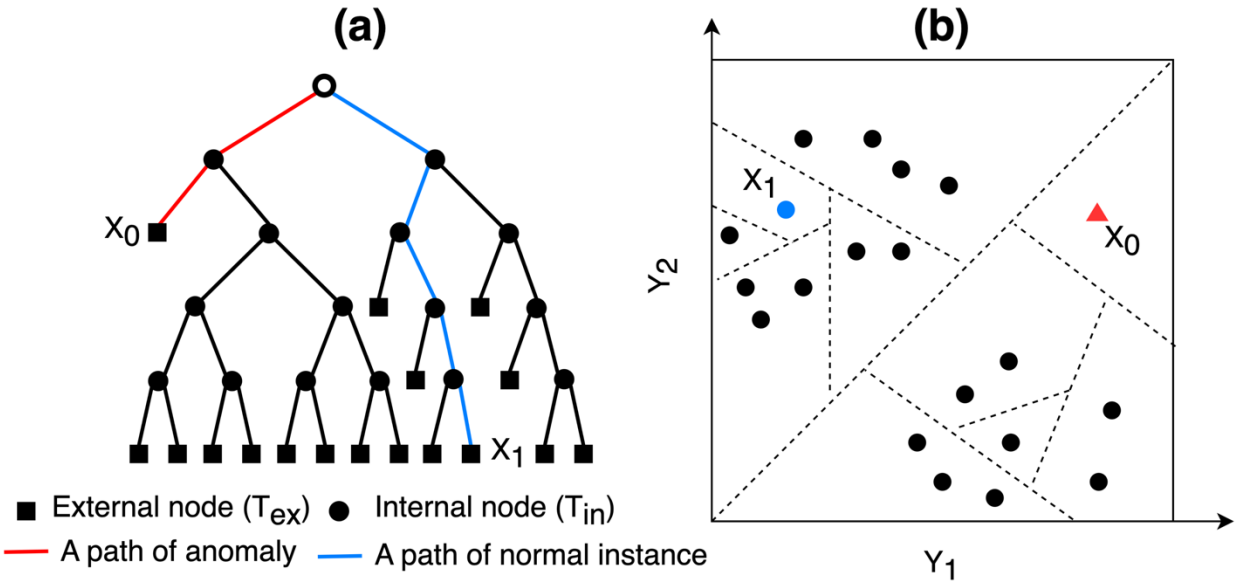

Figure 2-1. Schematic representation of a single tree (a) and its feature space (b) for an Extended Isolation Forest (EIF) built by a two-dimensional dataset.

After the whole dataset is split into trees, an anomaly score is calculated as follows:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{2-2}$$

Where $h(x)$ is the path length for external node terminations in one tree, and $E(h(x)$ is the mean $h(x)$ of all trees. $c(n)$ is the normalizing factor (Preiss, 2008). The calculated anomaly score s ranges from 0 to 1: the closer to 1, the more likely the sample is anomalous; otherwise, the

sample is more likely to be normal. To fit iForest into the SDM workflow, *itsdm* uses a linear conversion ($p = 1 - s$) to translate the anomaly score (s) into environmental suitability (p).

As a popular algorithm in anomaly detection, iForest has been continuously improved with amended node splitting methods such as split-criterion iForest (Liu et al., 2010), Robust random cut forest (Guha et al., 2016), and Fair-cut forest (Cortes, 2021b) , as well as new metrics for calculating outlier scores (Cortes, 2021a). The R (R Core Team, 2021) package *isotree* (Cortes, 2022) is an ensemble of iForest and these variants with fast and multi-threaded implementation and thus is used in *itsdm* for model training. Table B-1 (4.5Appendix B) lists the decisive arguments used in the function *isotree_po* (Table 2-1) for specific model types in the family of iForest.

2.5 Application of Shapley values

2.5.1 Local explanation and applications in *itsdm*

The Shapley value (Shapley, 1953) is an idea from cooperative game theory, which fairly distributes a game's payouts among players. The SHapley Additive exPlanations (SHAP) is an additive feature attribution method based on Shapley values that decomposes individual predictions of a model into the sum of the contributions of each variable value (Lundberg & Lee, 2017). Assume there is a prediction $f(x)$ for a single input x, the additive feature attribution method specifies the explanation as (Lundberg & Lee, 2017):

$$g(x') = \emptyset_0 + \sum_{i=1}^{M} \emptyset_i x_i' \tag{2-3}$$

where g is the explanation model. x' is the simplified x that maps to the original x by function $x = h_x(x')$. M is the number of input features. $\emptyset_0$ is the constant value when all inputs are missing, and $\emptyset_i \in \mathbb{R}$ is the feature attribution for feature i. It was theoretically proved that Shapley values are the unique solution of Eq. 3 with three desirable properties (see details in Lundberg & Lee, 2017):

$$\emptyset_i = \sum_{Q \subseteq S \backslash \{i\}} \frac{|Q|!\,(|S| - |Q| - 1)!}{|S|!} [f_{Q \cup \{i\}}(x_{Q \cup \{i\}}) - f_Q(x_Q)] \tag{2-4}$$

where S is the set of all features in the model. Q is a subset of S. $f_{Q \cup \{i\}}$ is a model trained with feature i present and f_Q is a model trained with feature i withheld. Thus, $f_{Q \cup \{i\}}(x_{Q \cup \{i\}}) - f_Q(x_Q)$ represents the effects of including feature i on the model. Because the effect of withholding a feature relies on other features, $\emptyset_i$ calculates the weighted average of $f_{Q \cup \{i\}}(x_{Q \cup \{i\}}) - f_Q(x_Q)$ of all possible subsets $Q \subseteq S \backslash \{i\}$.

Several approaches (e.g., Kernel SHAP and Linear SHAP) (Molnar, 2020; Štrumbelj & Kononenko, 2014) have been proposed to approximate Shapley values (2-4). The package *fastshap* (Greenwell, 2021) is used in *itsdm* to estimate Shapley values, in which a Monte-Carlo sampling approach (Štrumbelj & Kononenko, 2014) is efficiently implemented.

Shapely values demonstrate how each explanatory covariate pushes the model result from the base value (the average model output over the training dataset) (Molnar, 2020). Positive values vote for presence, and negative values vote for absence. The higher the absolute Shapley value is, the more important the explanatory variable is. Using Shapley values and the characteristics, function *variable_contrib* (Table 2-1) in *itsdm* can diagnose how the explanatory variables decide the environmental suitability at each observation location.

2.5.2 Global explanation and applications in *itsdm*

Additionally, Shapley values can be integrated into global explanations such as variable importance and response curves. Because features with large absolute Shapley values are important, variable importance could be evaluated by averaging the absolute Shapley values per feature across the whole data:

$$I_i = \frac{\sum_{j=1}^{n} \left| \emptyset_i^{(j)} \right|}{n} \tag{2-5}$$

This is implemented in function *variable_analysis* (Table 2-1) in package *itsdm*.

Shapley values technique shows how a species responds to an environmental variable by plotting all possible feature values $\{x_i^{(j)}\}_{j=1}^{n}$ against the corresponding Shapley values $\{\emptyset_i^{(j)}\}_{j=1}^{n}$. As Shapley values are signed, the response curves also can show the tipping point(s) of when this species starts to be negatively impacted by this environmental variable. In *itsdm*, *shap_dependence* (Table 2-1) is the function to generate Shapley values-based response curves. To illustrate how a variable affects prediction spatially, *itsdm* provides the function *shap_spatial_response* (Table 2-1), which uses Shapley values to generate spatial response maps.

Expanding from *shap_dependence* and *shap_spatial_response*, *itsdm* provides a unique function (*detect_envi_change*) to analyze the vulnerable areas potentially impacted by the changing environmental variables. The users can apply a number to the current environmental variable or assign a completely new future environmental variable. As a model-agnostic post-hoc method (Ryo et al., 2021), Shapley values technique can be used to explain any predictive models, therefore, the Shapley values-based functions in *itsdm* including *detect_envi_change*,

can apply to any SDMs. For instance, we fitted a MaxEnt SDM (Phillips & Dudík, 2008) named *mod_maxent* with multiple Bioclimatic variables BIO1, BIO2, BIO3, BIO13, BIO14, BIO18, and BIO19 (Fick & Hijmans, 2017) to estimate the habitat suitability of Za Baobab tree (*Adansonia za Baill.*) in Madagascar (see 'Data availability' section for code). With a *stars* (Pebesma, 2022) object called *bios_current* to represent the current environment, a *stars* (Pebesma, 2022) object called *bios_future* to represent the future (2041-2060) environment (Fick & Hijmans, 2017), and a wrapper function called *pfun* for *mod_maxent* to do prediction, *detect_envi_change* works as follows to detect potential impacts to Za Baobab tree by a changing BIO1 (annual mean temperature):

```
pfun <- function(X.model, newdata) {
  predict(X.model, newdata,
              args = c("outputformat=cloglog"))}
bio1_changes <- detect_envi_change(
  model = mod_maxent,
  var_occ = training[, 2:ncol(training)],
  variables = bios_current,
  target_var = "bio1",
  variables_future = bios_future,
  pfun = pfun)
```

The function returns a response curve with detected tipping points (Figure 2-2A), a vector of detected tipping points, a map of contribution change (Figure 2-2B), and a *stars* (Pebesma, 2022) object of contribution change.

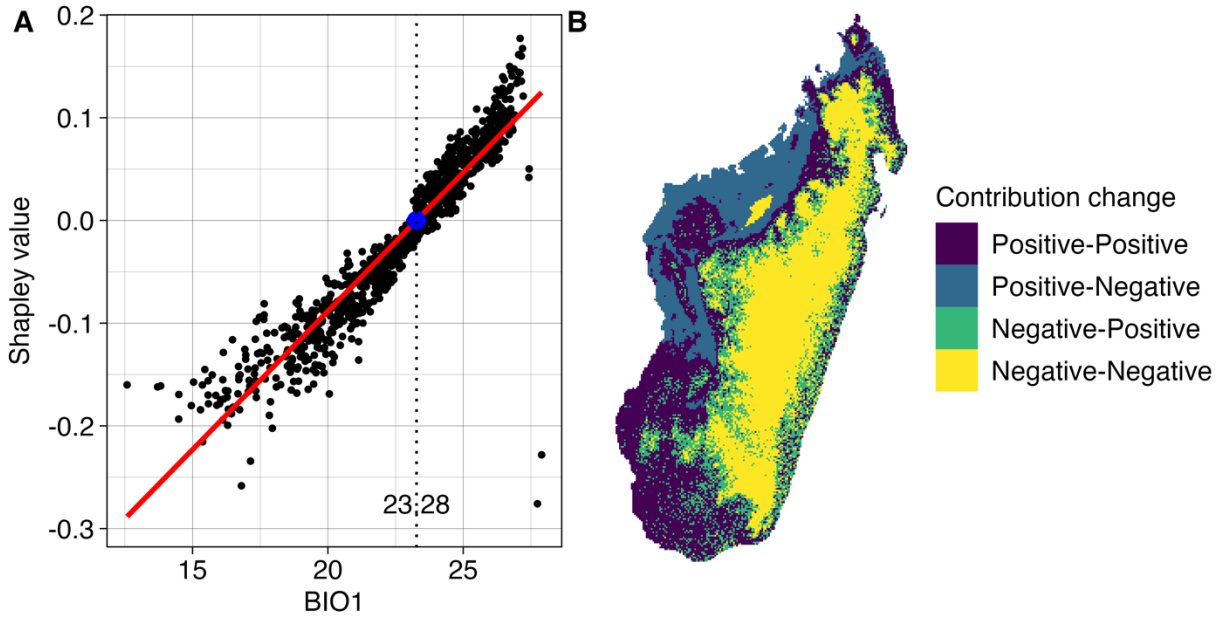

Figure 2-2. Environmental change analysis of BIO1 (annual mean temperature) to Za Baobab tree (*Adansonia za Baill.*) in Madagascar. Panel A shows that Za Baobab tree has a positive linear response to annual mean temperature in Madagascar. 23.28℃ is the tipping point of annual mean temperature, which means the Za Baobab tree in areas with an annual mean temperature near 23.28℃ is vulnerable to a cooling temperature. Panel B shows Za Baobab tree in most areas of Madagascar will not be affected by a changing annual mean temperature. The annual mean temperature in Northwest coastal areas will become not suitable for Za Baobab tree.

2.6 Example

To demonstrate the package functionality, we provide a short example using a virtual species generated by the package *virtualspecies* (Leroy et al., 2016), the distribution of which is in mainland Africa and shaped by climatic variables bio1, bio5, and bio12 (Fick & Hijmans, 2017). For this example, we took 2000 random presence-only samples and selected bio1, bio5, bio12, and three other unrelated features (var1 through var3) as the explanatory environmental variables. In the workflow, 70% of the samples are used for training, and 30% of them are used for evaluation. The details of the virtual species can be found in 4.5B.1.2 in 4.5Appendix B.

In function *isotree_po*, a model is fit to the provided *sf* (Pebesma, 2018) object of occurrence points and corresponding environmental variables, along with an optional *sf* (Pebesma, 2018) object of occurrence points for independent evaluation. For example, with a training set of occurrence points *obs*, an independent evaluation set called *eval*, and a *stars* (Pebesma, 2022) object holding the environmental predictors (*env_vars*), the following workflow creates an EIF model with an extension level of 2 (see more options in Table B-1) and a sampling rate of 0.8:

```
# Create an Extended isolation forest
mod <- isotree_po(
  obs = obs,
  obs_ind_eval = eval,
  variables = env_vars,
  sample_size = 0.8,
  ndim = 2)
```

The function *isotree_po* provides a highly automatic workflow that contains model creation, model evaluation, model prediction, and model explanation, with corresponding *print* and/or *plot* options to check the results (Table 2-1). The full description of the results can be found in section 4.5B.1.2 of 4.5Appendix B. Here we only present the Shapley values-based analysis.

If argument *check_variable* is set to *FALSE* in function *isotree_po*, the users can call function *variable_analysis* to diagnose variable importance:

```
var_analysis <- variable_analysis(
  model = mod$model,
  pts_occ = mod$observation,
  pts_occ_test = mod$independent_test,
  variables = mod$variables)
plot(var_analysis)
```

The function ranks environmental variables based on Shapley values (Figure 2-3) as well as the leave-one-out Jackknife test (section 4.5B.1.2 and Figure B-5 in Appendix 4.5B.1).

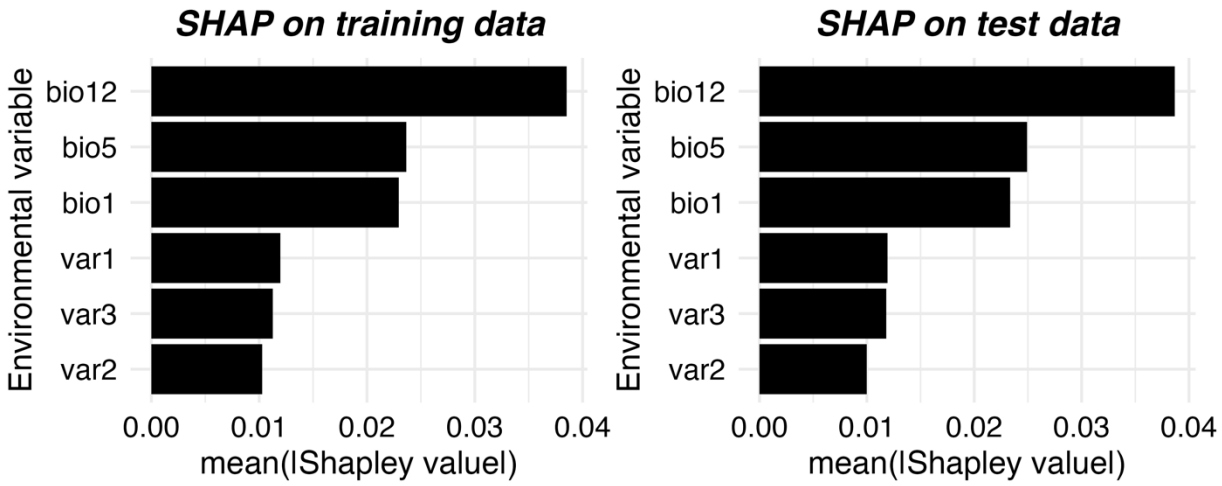

Figure 2-3. Variable importance of virtual species case diagnosed by Shapley values technique. Variables bio12, bio5, and bio1 have much higher importance than var 1 through var 3, as intended. In addition, the similarity in the values for these metrics for both the training and test dataset indicates that the model is generalizable.

itsdm employs several methods to generate response curves, including spatial ones. The Shapley value-based response curve conveys how prediction is pushed away from the average prediction across the whole training dataset (section 2.5). The Shapley values also allow users to diagnose the correlation between two variables. For example, Shapley value-based response curves of bio1 and bio12 are plotted and colored by bio5 (Figure 2-4):

```
# Plot Shapley value-based response curves without smoothing
plot(mod$shap_dependences,
  target_var = c('bio1', 'bio12'),
  related_var = 'bio5', smooth_line = FALSE)
```

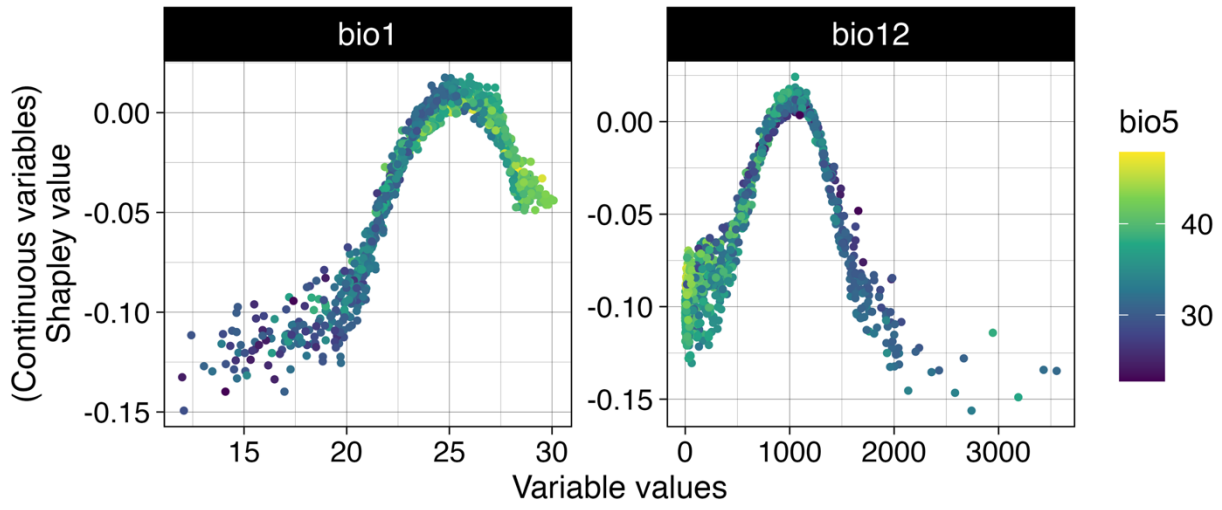

Figure 2-4. Shapley value-based response curves of bio1 and bio12 colored by bio5 in our virtual species case. The modeled species has a strong positive response to both bio1 and bio12 that respectively peak at 25 °C and 1000 mm, and that the two are also strongly correlated with bio5, particularly in the upper range for bio1 and in the lower to mid-range for bio12.

It is recommended to use response curves together with variable importance analysis to explain model inputs. However, the standard response curves only provide a graphical, non-spatial assessment of how a variable influences prediction. To illustrate how a variable affects prediction spatially, *itsdm* provides the function *spatial_response*, which generates spatial response maps. To calculate response maps, *spatial_response* is used with a non-zero *shap_nsim*:

```
# Make spatial response maps with all three methods
# Make sure to set a non-zero shap_nsim
full_spatial_responses <- spatial_response(
  model = mod$model,
  var_occ = mod$vars_train,
  variables = mod$variables,
  shap_nsim = 10)
plot(full_spatial_responses, target_var = 'bio12')
```

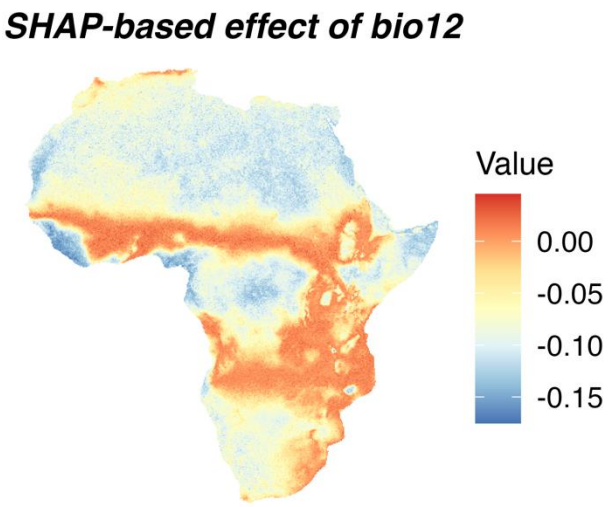

Figure 2-5. Shapley values-based spatial response map of variable bio12 in our virtual species case. It is evident that bio12 contributes minimally in some areas even though it is the most vital environmental variable diagnosed in variable analysis.

The last line displays the spatial response maps of variable bio12, and the Shapley values-based one is shown in Figure 2-5. Areas with Shapley values below zero are where bio12 votes for absence for this species, and areas with Shapley values above zero are the opposite. Variables with large absolute Shapley values contribute more than others (section 2.5).

itsdm also includes several optional post-analysis steps (Table 2-1), such as analyzing variable contributions to a specific observation, as shown in Figure 2-6. The figure shows that bio5 pushes the predicted suitability higher and votes for presence. The bio12 and bio 1 contribute oppositely.

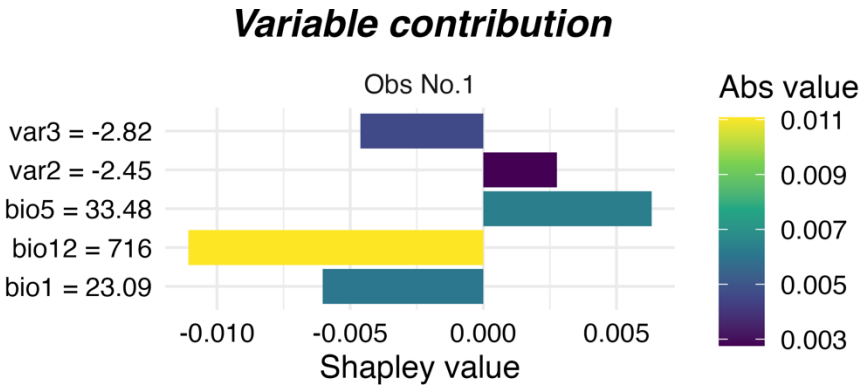

Figure 2-6. Variable contributions to the modeled suitability of an occurrence observation.

The full list of functions and additional examples can be found in the *itsdm* package documentation and package vignettes with extended examples.

2.7 Comparison with other SDMs and recommendations

To compare the predictive performance of iForest with other SDMs and highlight the conditions when it is beneficial to use iForest, we generated 50 virtual species with package *virtualspecies* (Leroy et al., 2016) (see 'Data availability' section for code). Bioclimatic variables BIO1, BIO2, BIO5, BIO6, BIO12, and BIO15 (Fick & Hijmans, 2017) in mainland Africa were used to simulate these species. We generated a species suitability map by applying a Gaussian, linear, logistic, or quadratic function with random parameters on randomly selected 3-5 variables for each species (Leroy et al., 2016). The final suitability is a multiplicative function of responses to the selected variables. A threshold of 0.5 or 0.6 was used to convert suitability to presence-absence to represent the normal detection type (Figure 2-7). A threshold of 0.8 or 0.9 was used to represent the core area concentrated detection type (Figure 2-7). A prevalence-weighted random number from 100 to 500 of presence-only samples were drawn from presence-absence map for

both detection types. In addition, 10000 background samples were taken for all SDMs except iForest. For evaluation, 2000 presence-absence samples were drawn for each species by excluding all training presence locations and their 3×3 neighbors and then subset the majority class to ensure class balance.

True skill statistics with a threshold of 0.5 ($TSS_{0.5}$) and three threshold-independent evaluation metrics: Area under the ROC curve (AUC), Pearson correlation (COR), and Euclidean distance, were used to assess predictive performance. AUC and $TSS_{0.5}$ measure the capability of a model to separate presences from absences. COR values and Euclidean distances in this experiment were calculated between the predicted environmental suitability and the simulated suitability of the virtual species. They work together to measure the similarity between predicted and actual suitability values, which is to say, they have the same values for the same cases.

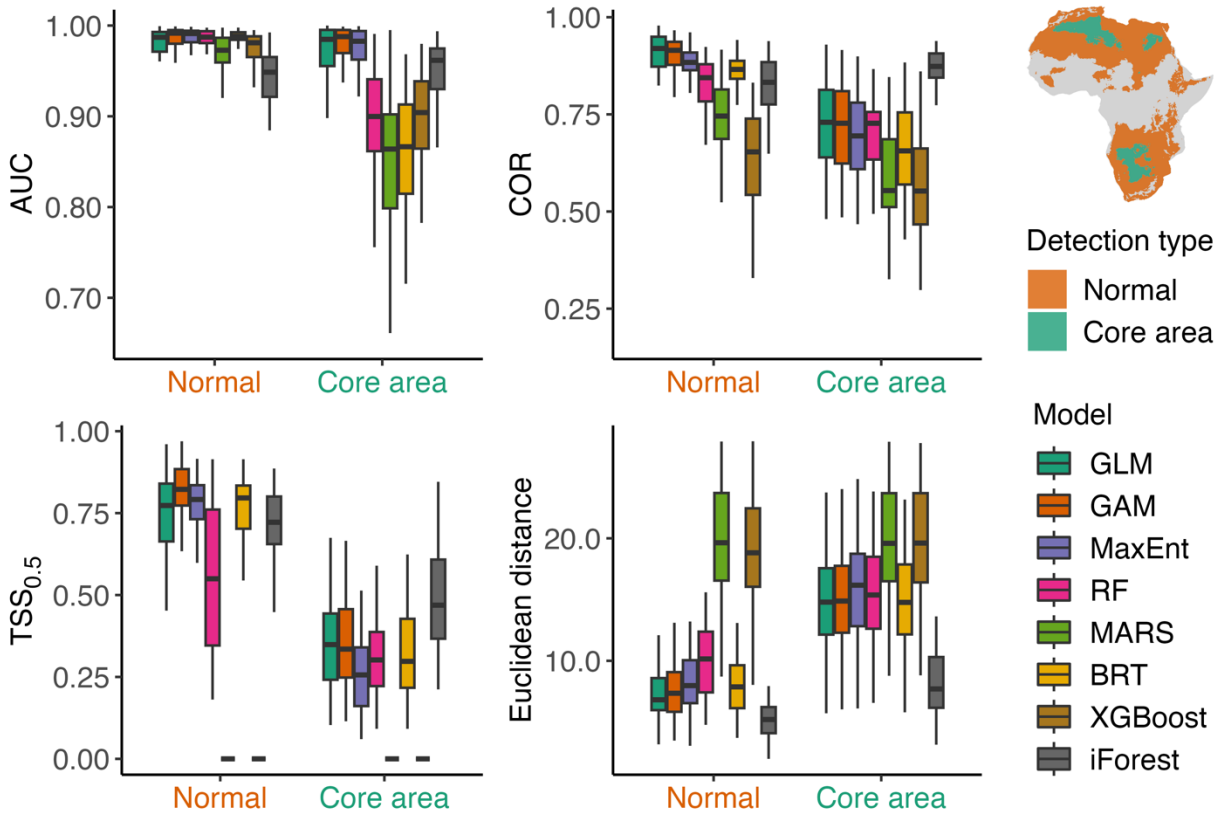

Figure 2-7. Performance comparison between Isolation Forest (iForest) and other mainstream SDM models. Evaluation metrics are Area Under the ROC Curve (AUC) , Pearson correlation between modeled and real habitat suitability (COR), True Skill Statistic with a threshold of 0.5 ($TSS_{0.5}$), and Euclidean distance between modeled and real suitability. Normal detection type is the case that the species can be detected in any areas with suitability higher than 0.5 or 0.6 and Core area type is the case that the species can only be detected in areas with suitability higher than 0.8 or 0.9. The figure in upright is an example drawn with No.6 virtual species.

We selected 7 SDMs with high-performance (Valavi et al., 2022) to make the comparison: Generalized linear model (GLM), Generalized additive model (GAM), Maximum entropy (MaxEnt), Random forest (RF), multivariate adaptive regression spline (MARS), Boosted regress trees (BRT), and Extreme gradient boosting (XGBoost). The results are shown in Figure 2-7. If the training samples can represent the actual distribution well (Normal case in Figure 2-7), GAM and GLM perform better than iForest and others, having a greater ability to discriminate presences and absences (high AUC and $TSS_{0.5}$) and higher similarity to actual

suitability (high COR and Euclidean distance). It is worth mentioning that suitability values predicted by iForest have the closest Euclidean distance with actual suitability values, which is also evident in Figure B-1. If the training samples only represent the actual distribution partially, e.g. having sampling bias or imperfect detection (Core area case in Figure 2-7), iForest starts to be advantageous. Even though iForest gets slightly lower AUC than GLM, GAM, and MaxEnt, it makes significantly higher COR and $TSS_{0.5}$ and lower Euclidean distance. The similar performance of models fitted under two cases (Figure 2-7 & Figure B-1) indicates that iForest is resistant to sampling issues and overfitting.

For presence-only species distribution modeling, when the species occurrences cover the true distribution range, models like GLM, GAM, or MaxEnt perform better than iForest. This is also true if occurrences do not cover the whole distribution range, but the range is known so that background samples can be extracted conditionally. When it is unfeasible to estimate the distribution range before modeling, iForest can be a cautious choice.

2.8 Discussion

iForest is an appealing method in SDM because it takes presence-only data as input, which matches it with most occurrence datasets of wildlife nowadays. Additionally, splitting feature space by hyperplanes is similar to profile models that translate species-environment relationships into profiles. Thus, it results in environmental suitability rather than the probability of presence. Unlike methods that are optimized to suitable conditions, iForest is optimized to describe unsuitable conditions and thus is less likely to overfit (Abe et al., 2006; He et al., 2003; Rousseeuw & van Driessen, 1999). These give iForest strengths as an SDM, particularly when it

is unclear if the presence samples cover suitable areas fully and there are no reliable absence samples to use.

Shapley values technique is a growing topic of interest in interpretable machine learning, as they can help to explain any predictive model (Ryo et al., 2021). It offers a potentially powerful tool to comparably interpret SDMs that are built with different methods and decipher complex models to explain real-world ecological phenomena (Mammola et al., 2019). As a post-hoc technique, Shapley values can be used to interpret the impacts of a changing environment in species distribution conveniently.

The R package *itsdm* offers convenient functions to fit iForest SDM and generalizes the Shapley values technique for all SDMs to analyze species' response to the environment. Undoubtedly, not relying on causal mechanisms, iForest and Shapley values technique have the same limitations as other statistical methods in applications of ecological modeling, especially for change analysis. *itsdm* is intended as a new SDM toolbox that complements existing frameworks, which will enable users to apply iForest and Shapley values technique in their studies and explore advantages and disadvantages.

Acknowledgements

This project was supported by the Future Investigators in NASA Earth and Space Science and Technology (FINESST) program (award number: 80NSSC20K1640). The authors thank David Cortes for the suggestion of improving the code flexibility and the authors for all the fabulous and valuable packages that *itsdm* depends on.

Conflict of interests

The authors have no conflict of interest.

Authors' contributions

Lei Song conceived the ideas, collected example data, and analyzed the data; Lei Song and Lyndon D. Estes designed methodology and led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Data accessibility statement

The *itsdm* package, documentation, and example data are hosted and available on CRAN (https://cran.r-project.org/package=itsdm). The source code can be assessed at GitHub (https://github.com/LLeiSong/itsdm), and version 0.2.0 of the package used for this manuscript is archived on Zenodo (Song & Estes, 2023). All simulation species and scripts not shown in supplementary materials are available via Open Science Framework (OSF): https://osf.io/8mc4e/.

ORCID

Lei Song: https://orcid.org/0000-0002-4371-1473

Lyndon D. Estes: https://orcid.org/0000-0002-9358-816X

References

Abe, N., Zadrozny, B., & Langford, J. (2006). Outlier detection by active learning. *In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 504–509. https://doi.org/10.1145/1150402.1150459

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, *3*(2), 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x

Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, *19*, 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002

Cortes, D. (2021a). Isolation forests: Looking beyond tree depth. *ArXiv:2111.11639 [Cs, Stat]*. http://arxiv.org/abs/2111.11639

Cortes, D. (2021b). Revisiting randomized choices in isolation forests. *ArXiv:2110.13402 [Cs, Stat]*. http://arxiv.org/abs/2110.13402

Cortes, D. (2022). *isotree: Isolation-Based Outlier Detection*. https://CRAN.R-project.org/package=isotree

Elith, J., Ferrier, S., Huettmann, F., & Leathwick, J. (2005). The evaluation strip: A new and robust method for plotting predicted responses from species distribution models. *Ecological Modelling*, *186*(3), 280–289. https://doi.org/10.1016/j.ecolmodel.2004.12.007

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17*(1), 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x

Feremans, L., Vercruyssen, V., Cule, B., Meert, W., & Goethals, B. (2020). Pattern-based anomaly detection in mixed-type time series. *Machine Learning and Knowledge Discovery in Databases*, 240–256.

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302–4315. https://doi.org/10.1002/joc.5086

Franklin, J. (2010). *Mapping Species Distributions: Spatial Inference and Prediction*. Cambridge University Press.

Greenwell, B. (2021). *fastshap: Fast Approximate Shapley Values*. https://CRAN.R-project.org/package=fastshap

Guha, S., Mishra, N., Roy, G., & Schrijvers, O. (2016). Robust random cut forest based anomaly detection on streams. *International Conference on Machine Learning*, 2712–2721.

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, *135*(2–3), 147–186. https://doi.org/10.1016/S0304-3800(00)00354-9

Hariri, S., Kind, M. C., & Brunner, R. J. (2019). Extended Isolation Forest. *IEEE Transactions on Knowledge and Data Engineering*, 1–1. https://doi.org/10.1109/TKDE.2019.2947676

He, Z., Xu, X., & Deng, S. (2003). Discovering cluster-based local outliers. *Pattern Recognition Letters*, *24*(9–10), 1641–1650. https://doi.org/10.1016/S0167-8655(03)00003-5

Khan, S., Liew, C. F., Yairi, T., & McWilliam, R. (2019). Unsupervised anomaly detection in unmanned aerial vehicles. *Applied Soft Computing*, *83*, 105650.

Kramer-Schadt, S., Niedballa, J., Pilgrim, J. D., Schröder, B., Lindenborn, J., Reinfelder, V., Stillfried, M., Heckmann, I., Scharf, A. K., Augeri, D. M., Cheyne, S. M., Hearn, A. J., Ross, J., Macdonald, D. W., Mathai, J., Eaton, J., Marshall, A. J., Semiadi, G., Rustam, R., … Wilting, A. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, *19*(11), 1366–1379. https://doi.org/10.1111/ddi.12096

Leroy, B., Meynard, C. N., Bellard, C., & Courchamp, F. (2016). Virtualspecies, an R package to generate virtual species distributions. *Ecography*, *39*(6), 599–607. https://doi.org/10.1111/ecog.01388

Li, S., Zhang, K., Duan, P., & Kang, X. (2019). Hyperspectral anomaly detection with kernel isolation forest. *IEEE Transactions on Geoscience and Remote Sensing*, *58*(1), 319–329.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. *2008 Eighth IEEE International Conference on Data Mining*, 413–422.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2010). On detecting clustered anomalies using SCiForest. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 274–290.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-Based Anomaly Detection. *ACM Transactions on Knowledge Discovery from Data*, *6*(1), 1–39. https://doi.org/10.1145/2133360.2133363

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777.

Mammola, S., Milano, F., Vignal, M., Andrieu, J., & Isaia, M. (2019). Associations between habitat quality, body size and reproductive fitness in the alpine endemic spider *Vesubia jugorum*. *Global Ecology and Biogeography*, *28*(9), 1325–1335. https://doi.org/10.1111/geb.12935

Merow, C., Smith, M. J., & Silander, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: What it does, and why inputs and settings matter. *Ecography*, *36*(10), 1058–1069. https://doi.org/10.1111/j.1600-0587.2013.07872.x

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com. https://christophm.github.io/interpretable-ml-book/

Morales, N. S., Fernández, I. C., & Baca-González, V. (2017). MaxEnt's parameter configuration and small samples: Are we paying attention to recommendations? A systematic review. *PeerJ*, *5*, e3093. https://doi.org/10.7717/peerj.3093

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, *10*(1), 439–446. https://doi.org/10.32614/RJ-2018-009

Pebesma, E. (2022). *stars: Spatiotemporal Arrays, Raster and Vector Data Cubes*. https://CRAN.R-project.org/package=stars

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3–4), 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with Maxent: New extensions and a comprehensive evaluation. *Ecography*, *31*(2), 161–175. https://doi.org/10.1111/j.0906-7590.2008.5203.x

Preiss, B. R. (2008). *Data structures and algorithms with object-oriented design patterns in C++*. John Wiley & Sons.

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Radosavljevic, A., & Anderson, R. P. (2014). Making better MAXENT models of species distributions: Complexity, overfitting and evaluation. *Journal of Biogeography*, *41*(4), 629–643. https://doi.org/10.1111/jbi.12227

Rousseeuw, P. J., & van Driessen, K. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, *41*(3), 212. https://doi.org/10.2307/1270566

Ryo, M., Angelov, B., Mammola, S., Kass, J. M., Benito, B. M., & Hartig, F. (2021). Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. *Ecography*, *44*(2), 199–205.

Shapley, L. S. (1953). A value for n-person games. In *Contributions to the Theory of Games* (2.28, pp. 307–317).

Sofaer, H. R., Jarnevich, C. S., Pearse, I. S., Smyth, R. L., Auer, S., Cook, G. L., Edwards, T. C., Guala, G. F., Howard, T. G., Morisette, J. T., & Hamilton, H. (2019). Development and Delivery of Species Distribution Models to Inform Decision-Making. *BioScience*, *69*(7), 544–557. https://doi.org/10.1093/biosci/biz045

Song, L., & Estes, L. (2023). *Itsdm* (v0.2.0). Zenodo. https://doi.org/10.5281/zenodo.7533022

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, *41*(3), 647–665. https://doi.org/10.1007/s10115-013-0679-x

Thuiller, W., Georges, D., Gueguen, M., Engler, R., & Breiner, F. (2021). *biomod2: Ensemble Platform for Species Distribution Modeling*. https://CRAN.R-project.org/package=biomod2

Thuiller, W., Lafourcade, B., Engler, R., & Araújo, M. B. (2009). BIOMOD - a platform for ensemble forecasting of species distributions. *Ecography*, *32*(3), 369–373. https://doi.org/10.1111/j.1600-0587.2008.05742.x

Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J., & Elith, J. (2022). Predictive performance of presence-only species distribution models: A benchmark study with reproducible code. *Ecological Monographs*, *92*(1). https://doi.org/10.1002/ecm.1486

Chapter 3

A national, multi-scale assessment of habitat connectivity of African savanna elephant (*Loxodonta africana*)

3.1 Abstract

Populations of African savanna elephants declined by an estimated 60% since 1970s, due to a combination of hunting and habitat loss, as human activities and environmental change have caused unprecedented habitat loss and fragmentation. Conserving habitat corridors for elephants is thus increasingly important for maintaining ecological and genetic connectivity, in order to conserve remaining populations. Using landscape connectivity modeling, we aimed to understand population connectivity and target vital corridors and hotspots across Tanzania, one of the most important elephant range states. We developed a multi-scale integrated species distribution model (SDM) using the Isolation Forest (iForest) algorithm to estimate spatial suitability of African savanna elephants. Two SDMs based on polygon-based observations (expert range map and census blocks) and presence-only occurrences were created separately and then ensembled by Bayes fusion. In particular, a set of landscape metrics was employed to describe the detailed landscape structure that was generated from a high-resolution land cover map with a resolution of 4.77m. We then used the environmental suitability to map landscape connectivity using circuit theory with Circuitscape. Our results indicate that both long-distance

and short-distance corridors are currently facing significant threats from intensive human activities, and our analysis also reveals a significant risk of nationwide connectivity reduction if agriculture and settlement continue to sprawl in Tanzania. In addition, our study identifies priority corridors and hotspots that should be targeted for connectivity conservation efforts to maintain population viability.

KEYWORDS: African savanna elephant, Isolation Forest, multi-scale, movement model, circuit theory, habitat fragmentation, human impact

3.2 Introduction

The global network of protected areas (PAs) has significantly expanded over the past 50 years and currently covers around 15% of the terrestrial surface (Geldmann et al., 2019), however, establishing isolated PAs may not be sufficient for biodiversity preservation (Berger, 2004; Newmark, 2008). Previous studies have shown that many PAs were designed in the absence of sufficient ecological forethought, often without considering connectedness, the increased disturbances outside PAs, and the movements of large mammals (Bleich, 2016), leading to the isolation of PAs (Maxwell et al., 2020; Newmark, 2008; Saura et al., 2017). Maintaining ecological connectivity is particularly challenging in rapidly developing regions with high rates of agricultural transformation and human population increase (Loos, 2021; Mammides, 2020), which is often the case in many African countries that also support large mammal populations with large geographic ranges (Bowyer et al., 2019; Raven et al., 2020). As human-driven biophysical disturbances have largely fragmented landscapes and disrupted the

connections between PAs, existing networks of PAs in these regions may be ineffective to safeguard large mammals, particularly for long-distance migration (Newmark, 2008). However, the ability to maintain free and wide-ranging movement of these megafauna over large landscapes is critical for sustaining individual species (e.g. genetic diversity and adaptation to climate change) as well as broader ecological functioning (e.g. seed dispersal and nutrient cycling), especially in the face of rapid climate and environmental changes (Berger, 2004; Maxwell et al., 2020). As one of the most important migratory megafauna, African savanna elephant (*Loxodonta africana*) play a critical role in structuring natural ecosystems and supporting a broad range of other species (Kohi et al., 2011). For instance, elephants manipulate woody vegetation structure in savannas by affecting fire regimes and species composition (Goheen et al., 2010; Pringle et al., 2016). The seasonal migration of elephants can facilitate habitat restoration and alleviate pressure on habitats for resources (Gara et al., 2021). By traveling over vast landscapes and dispersing seeds in their dung, elephants can help maintain corridors for other animals and promote recruitment and re-establishment of vegetation (Vidal et al., 2013).

Nonetheless, following population declines over several decades, they were recently listed as Endangered on the International Union for Conservation of Nature (IUCN) Red List of Threatened Species (Gobush et al., 2022). The population of African savanna elephants has dropped by more than 60% since the 1970s, primarily caused by the poaching crisis (Douglas-Hamilton, 1987; Wittemyer et al., 2014). Particularly, in Eastern Africa, they experienced an approximately 50% decline since 2007, mainly due to the considerable population losses in Tanzania (Thouless et al., 2016). It is noteworthy that other underlying threats, such as habitat conversion, habitat fragmentation, and human-wildlife conflicts, are of critical conservation

importance and recently have received rising attention in conservative decision-making (Mpakairi et al., 2019; Newmark, 2008; Thouless et al., 2016). Multiple studies have found that the connectivity of elephant populations is under critical threat or has been lost due to environmental change and human impacts (Lohay et al., 2020; Newmark, 2008). The isolation of habitats decreases their effective size to make them too small to be sustainable (Newmark, 2008). Moreover, disrupting the movement of elephants between and within ecosystems limits the gene flow between the populations, threatening their genetic and evolutionary sustainability and long-term population viability (Macdonald et al., 2013; Wall et al., 2013). Meanwhile, the competition between land suitable for agriculture and migratory corridors can lead to increased human-elephant conflicts (Green et al., 2018).

To better protect elephants or any other large mammals from these rising pressures they face, current conservation management requires and starts to investigate and take strategies (e.g. identify important migration corridors) to maintain and restore the linkage between PAs (Saura et al., 2019; Tshipa et al., 2017). For example, in Kenya, creation of new corridors is now included in the conservation targets (Green et al., 2018). Despite their potential contribution to policy making, existing studies on habitat connectivity have unneglectable limitations. Studies that identify connections of wildlife populations by accounting for spatially explicit resistance to animal movement typically focus on small scales, such as national parks (Bastille-Rousseau & Wittemyer, 2021; Bukombe et al., 2022; Epps et al., 2013; Gara et al., 2021; Osipova et al., 2019a, 2019b), because monitoring animal movement across vast regions is both logistically challenging and expensive (Tshipa et al., 2017). These studies are suitable for local management. Other studies that evaluate connectivity between PAs on large scales (e.g. global) often rely solely on the geometric distance between the areas and do not consider the permeability of the

land between them (Saura et al., 2017, 2019; Ward et al., 2020). This assessment is unrealistic for many regions, especially those that are highly disturbed by human activities.

In this study, we used an approach that combines multi-scale species distribution modeling and graph-based landscape connectivity modeling to identify corridors of elephants within and between habitat blocks throughout Tanzania. Particularly, we used a set of landscape metrics derived from a very high-resolution (4.77 m) land cover map to model environmental suitability, which was used as movement conductance matrix to evaluate spatially explicit landscape connectivity. By overlapping human disturbances, we also identify corridors at greatest risk of loss due to habitat conversion. Our research provides new information that can be used to help inform broader efforts within a critical elephant range state to balance development and conservation objectives.

3.3 Materials and methods

3.3.1 Study area

The United Republic of Tanzania (Figure 3-1), located in East Africa, is a mountainous country with diverse topography and climate caused by the western and eastern branches of the East African Rift System (EARS) (Rowhani et al., 2011). This has led to the formation of various ecosystems such as savanna, tropical and subtropical forests, and montane, which offer abundant natural resources (Tanzania National Bureau of Statistics, 2021). Tanzania thus boasts an impressive level of biodiversity, home to more than 300 mammal species, including 20% of the large mammal species found in Africa, as well as numerous birds, amphibians, and reptiles (Nkwabi et al., 2018). Particularly, Tanzania is renowned for its large elephant population, and it

contains several of the largest elephant habitats, including the Serengeti National Park, Tarangire National Park, and Selous Game Reserve, as well as significant elephant migration routes. Despite the fact that over 30% of the country's territory is designated as protected areas (PAs), Tanzania has faced a significant decline in its elephant population, dropping by over 60% between 2009 and 2014, primarily as a result of poaching and habitat loss and fragmentation (Thouless et al., 2016). It also has the most threatened biodiversity on the African continent, with 1591 Critically Endangered, Endangered, and Vulnerable species listed as of 2023 (IUCN, 2022, pp. 2022–2).

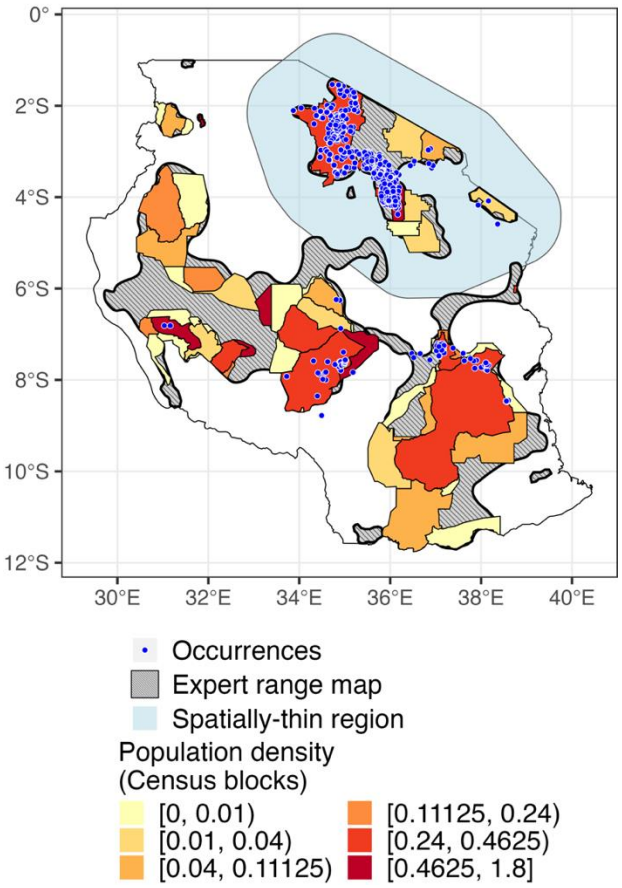

Figure 3-1. The study area and population observations (Expert range map, census blocks, and occurrences obtained from GBIF database) of African savanna elephants. Spatially-thin region delineates the region with overdense biased occurrences that need to be spatially thinned. (See details in section 3.3.3.2).

3.3.2 African savanna elephant distribution data

Two data sources were used as African savanna elephant location distribution references: polygon-based maps and point-based occurrences (Figure 3-1). The polygon-based data were obtained from the African Elephant Database (AfricanElephantDatabase.org) and consist of the International Union for Conservation of Nature (IUCN) range map and census blocks that define the input zones for elephant population censuses. Specifically, 13 polygons in Tanzania were identified as the known or possible ranges of African savanna elephants. The census dataset, organized by the African Elephant Specialist Group (AfESG), includes estimated population and the boundaries of 64 census survey zones in Tanzania after 2010, as provided by wildlife management agencies and other organizations (Thouless et al., 2016). In order to construct a species distribution model, several pseudo samples (Alhajeri & Fourcade, 2019; Fourcade, 2016; Rotenberry & Balasubramaniam, 2020) were drawn from both the expert range map and the census blocks, using population as a weight for sampling. Since the expert range map lacks population information, we computed a population value by calculating the average population density from all census blocks (0.18 elephants/km^2) and multiplying it by the range area and assign this value to each polygon fragments in the expert range map that is not covered by census blocks.

We obtained the coordinates of occurrences for the African savanna elephant in GBIF (http://www.gbif.org). To retrieve these occurrences, we utilized the *occ_search* function from the R package *rgbif* (Chamberlain et al., 2022; Chamberlain & Boettiger, 2017) and set the start year as 2015 and Tanzania as the search country. By setting arguments in function *occ_search*, we only retained occurrences that had valid coordinates and no geospatial issues to be used for

modeling. Additionally, another occurrence dataset utilized in our study was shared by Lohay et al. (2020) for their analysis on the genetic connectivity of African savanna elephants. This dataset consisted of 688 individual elephants sampled in 2015 and 2017 across four ecosystems that harbor the largest elephant populations in Tanzania.

The polygon-based dataset offers a more accurate representation of the spatial distribution of elephants, albeit with a lower location precision and higher commission errors (Brooks et al., 2019). On the other hand, the point-based dataset offers the actual occurrence of elephants, but has limited availability and a highly biased spatial distribution. In light of these limitations, we employed an integrated approach in this study by combining these two types of observations. This allowed us to mitigate their individual shortcomings and minimize the potential for biased predictions. The workflow for this approach is outlined in the following section 3.3.3.

3.3.3 Integrated multi-scale species distribution modeling

To map the environmental suitability of African savanna elephants in Tanzania, we employed an integrated multi-scale species distribution modeling approach that is illustrated in Figure 3-2. This modeling approach consists of three parts: regional scale species distribution modeling (SDM), landscape scale SDM, and their integration (Figure 3-2). In brief, we used polygon-based observations and environmental variables affecting elephant distribution at broad scales to develop a regional scale distribution model. Then, we used point-based observations and environmental variables constraining elephant occurrence at a fine scale to fit a landscape scale distribution model. Finally, we integrated the suitability predictions made by these two

models to obtain the final suitability prediction. The integrated method jointly considers regional and landscape drivers of elephant distributions and provides a strategy to combine multiple data sources that are considerably different in design and accuracy, for instance, the expert range map and occurrence survey dataset in this study (section 3.3.2).

Both regional and landscape species distribution model (SDM) were generated using algorithm Isolation Forest (Liu et al., 2008) with the *itsdm* R package (Song & Estes, 2023). Like other SDMs, Isolation Forest creates habitat suitability maps using species occurrence data (including presence-only) and a set of habitat covariates. It is particularly well-suited for situations where it is uncertain whether the presence samples adequately cover suitable areas and there are no reliable absence samples available to utilize, which is a common challenge for many presence-only datasets (Song & Estes, 2023). Within *itsdm* package, a Shapley-value-based approach (Lundberg & Lee, 2017; Molnar, 2020; Shapley, 1953) was used to analyze variable response and evaluate relative importance of environmental variables. The values possess directions, where positive values indicate a contribution to the presence in SDM, while negative values imply a contribution to the. Higher absolute values indicate greater importance absence (see more details in section 4.5C.1.2 in 4.5Appendix C). We primarily used three evaluation metrics to assess our modeling results: AUC, AUC_{ratio}, and F-measure. The AUC_{ratio} is the area under the ROC_{ratio} curve, which plots plotting the proportion of correctly predicted presence against the proportion of presences falling above a range of thresholds against the proportion of cells of the whole area falling above the range of thresholds (Peterson et al., 2008). The full description of evaluation metrics can be found in section 4.5C.1.2 in 4.5Appendix C.

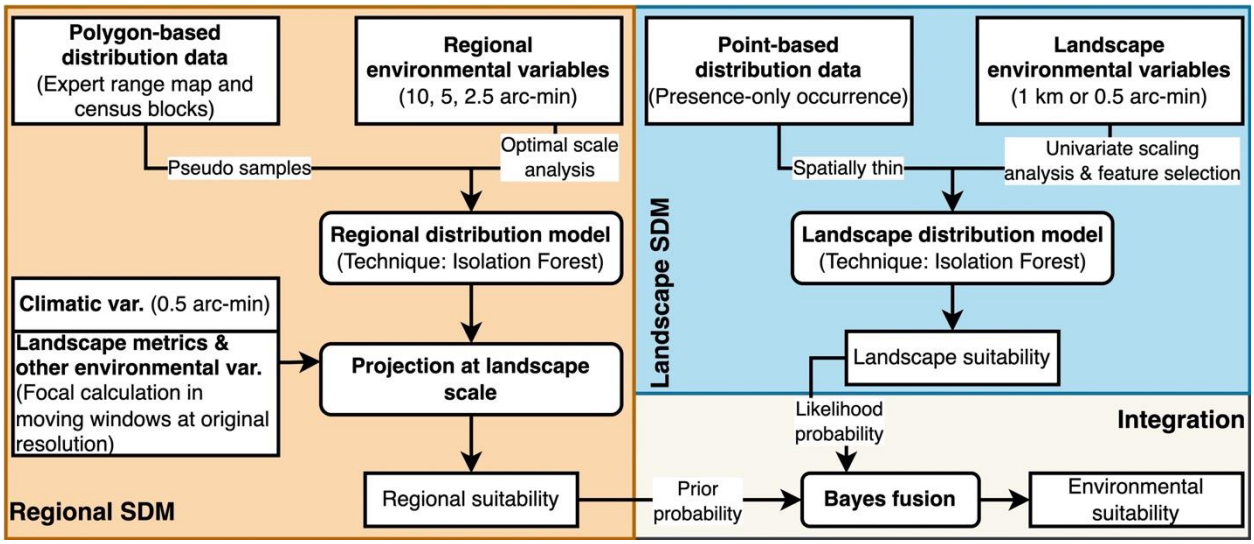

Figure 3-2. The integrated approach for multi-scale species distribution modeling. SDM means species distribution modeling. The var. is the abbreviation of variables.

Based on previous studies (Dejene et al., 2021; Ntukey et al., 2022; Williams et al., 2018), we initially identified three groups of variables that we hypothesized might influence the distribution of African savanna elephants in our study area. The primary covariates included vegetation (Normalized Difference Vegetation Index, NDVI), surface water availability, and three anthropogenic factors (roads, settlements, and agriculture), which represent food resources, water availability, and competition with human development. We also incorporated other environmental variables (climate, topographic factors, and landscape structure) that have the potential to influence elephant distribution. Uniquely, we utilized a land cover map with a spatial resolution of 4.77 m (Song et al., 2023) to provide detailed information on key landscape features (details provided in Text S1). To evaluate class-level landscape structure, we combined the areas classified as grassland, shrubland, and wetland, as they represent similar open habitats of African savanna elephants, and keep Forest/Dense tree as the closed canopy habitat (Table 3-1). Environmental variables were calculated in different ways for the regional and landscape scale SDMs and were tested independently for each SDM.

Detailed information on the modeling process and data utilized for each component of the integrated multi-scale species distribution modeling approach are explained in the following sections (3.3.3.1-3.3.3.3).

Table 3-1 Selected environmental variables for the regional and landscape species distribution model.

Category	Variable	Final usage	Original resolution[2]
Climate	Annual Mean Temperature (BIO1)	Regional	10, 5, 2.5, 0.5 arc-minutes
	Temperature Seasonality (BIO4)	Both	
	Temperature Annual Range (BIO7)	Regional	
	Annual Precipitation (BIO12)	Regional	
	Precipitation Seasonality (BIO15)	Regional	
Vegetation	NDVI (Wet season)	Both	30 meters
	NDVI (Dry season)	Regional	
	NDVI seasonality	Both	
Surface water sources	Coverage of waterbodies	Regional	4.77 meters
	Rivers	Both[1]	--
Antropologic factors	Settlements	Both[1]	--
	Cropland edge density	Both	4.77 meters
	Primary roads/railways	Both[1]	--
Topographic factors	Surface roughness	Both	30 meters
	Vector Ruggedness Measure (VRM)	Landscape	30 meters
Landscape metrics (Landscape level)	Patch density (PD)	Regional	4.77 meters
	Patch richness density (PRD)	Regional	
	Shannon's diversity index (SHDI)	Regional	
Landscape metrics (Class level)	Coverage of open habitat	Regional	4.77 meters
	Edge density (ED) of closed canopy habitat	Regional	
	Contiguity index distribution (CONTIG_MN) of open habitat	Both	
	Patch density (PD) of open habitat	Both	

[1]The density (length/area for rivers and roads, count/area for settlements) in each spatial grain is used for regional scale modeling, and the distance to features is used for landscape scale modeling (section 3.3.3).
[2]The climatic variables were used at their original resolutions, while the other variables were aggregated through zonal or focal calculation from their original resolutions to the target scales. The target scales differ between the regional and landscape SDMs (refer to section 3.3.3 and section 4.5C.1.2 in 4.5Appendix C), so even though they are listed under the same names, the variables are not the same.

3.3.3.1 Regional species distribution modeling

We chose three scales (2.5, 5, and 10 arc-minutes) to conduct the regional SDM and only kept the optimal one with best modeling performance (optimal scale analysis in Figure 3-2). To select the most relevant climate variables, we first examined the pairwise Spearman correlations (Hauke & Kossowski, 2011) between the 19 climate variables and picked five that

represent different climatic conditions and have correlations below 0.7 with other variables (Table 3-1 & more details in Table C-1). All other environmental variables listed in Table 3-1 were rescaled from their original resolution to target spatial scales using zonal calculation. We then used the pseudo samples generated from the polygon-based observations (as described in section 3.3.2) to construct SDMs at different scales. To represent the actual elephant population distribution, a sub-sampling strategy (an embedded feature of the algorithm) was used in the Isolation Forest model (Liu et al., 2008), with the population density of each census block that each pseudo-occurrence belongs to serving as sampling probabilities (Cortes, 2022).

The regional SDM with optimal scale was then projected at landscape scale (1 km or 0.5 arc-minutes, section 3.3.3.1) to calculate habitat suitability (Figure 3-2). Instead of being downscaled, all environmental variables (except climatic variables) used for projection were produced by focal calculation with a moving window of the same size as the spatial scale used in regional SDM. The resulting habitat suitability map represents the relationship between environment and elephant occurrence at a coarse scale described by the regional SDM but has a finer spatial resolution of 1 km.

3.3.3.2 Landscape species distribution modeling

To build the landscape-scale SDM, we relied on point-based occurrences with precise location, and implemented various strategies to mitigate the influence of sampling bias. The occurrences are highly concentrated in the northern area. In this area (as shown in the spatially-thin region in Figure 3-1), we used the R package *spThin* to reduce sampling bias and redundancy by ensuring that presence points were at least 20 km apart. The spThin approach randomly removes records that violate the minimum nearest neighbor distance (NND) constraint,

resulting in different thinned results under different randomization settings. To take advantage of this property, we spatially thinned the occurrences 10 times for cross-validation in this study, rather than thinning the occurrences once and splitting the result into 10 folds for cross-validation. This strategy enables us to avoid discarding valuable occurrence records and improve model generalization.

We selected the initial set of environmental variables based on variable analysis at regional SDM and a spatial resolution of 1 km for landscape SDM. These variables were produced by focal calculation on their original resolution using moving windows of different sizes (3, 5, and 7 km), except for the climate variable and vector-based features (listed in Table 3-1). We calculated distances to vector-based features to represent their influences at landscape scale. To increase their representative ability and reduce the impacts of sampling bias, we employed three criteria to select environmental variables. Due to the spatial bias and fragmentation of the real presence-only occurrences, they cannot fully represent the gradient of all environmental variables at the national scale. Therefore, we initially computed the Kolmogorov–Smirnov (K-S) distance (Massey Jr, 1951) between the values of variables at occurrence coordinates and the values of all variables across the expert range area. We excluded the variables with a K-S distance greater than 0.4 to ensure that the environmental variables can represent the conditions across suitable habitats in Tanzania (Figure C-1). Then, we conducted a univariate scaling (Sun et al., 2021; Timm et al., 2016) for each variable produced by focal calculation with different window sizes to identify the optimal window size that is most strongly related to African elephant occurrence. As the third step, we calculated the pairwise Spearman correlations (Hauke & Kossowski, 2011) on the remaining variables to avoid multicollinearity. We retained 9 variables whose correlation coefficient $|r|<0.7$ (Table 3-1).

After building Isolation Forest models with 10-fold cross-validation (as described earlier in this section), we conducted predictions and obtained a landscape suitability map by taking the pixel-wise average of the model outputs.

3.3.3.3 Bayes Theorem-based map integration

We integrated the species distribution models obtained at the coarse and fine scales using Bayes Theorem (Berrar, 2018; Drake & Richards, 2018), which is a mathematical formula used for calculating a conditional probability without the joint probability. The environmental suitability calculated by the coarse scale model (section **Error! Reference source not found.**) was viewed as the prior probability of occurrence $P(suit)$. The probability of occurrence obtained by the fine scale model (section **Error! Reference source not found.**) was taken as the likelihood probability of occurrence $P(occ|suit)$. The posterior probability of occurrence P_k at pixel k is calculated as follows (Berrar, 2018; Shen et al., 2021):

$$P_k = \frac{P_k(suit) \cdot P_k(occ|suit)}{P_k(suit) \cdot P_k(occ|suit) + (1 - P_k(suit)) \cdot (1 - P_k(occ|suit))} \quad (3\text{-}1)$$

3.3.4 Landscape connectivity

To assess the effects of human activities and locate at-risk corridors for maintaining connectivity among elephant populations, we utilized circuit theory using Circuitscape 5 (Hall et al., 2021) to analyze landscape connectivity. We divided Tanzania's elephant habitats into three primary clusters (Figure C-3 and section 3.3.4) and examined connectivity within and between these clusters. In this process, two scenarios were carried out to evaluate the current and potential impacts of human disturbances on landscape connectivity. In scenario A, we identified regions

with extensive human activities by selecting pixels (1 km^2) with over 80% cropland coverage (Figure C-2a) and more than 400 settlements (a dense town, Figure C-2b), and labeled them as barriers to elephant movement. In scenario B, we predicted regions that could potentially have intensive human activities by selecting pixels (1 km^2) currently with over 20% cropland coverage (Figure C-2c) and more than 200 settlements (double the current density, Figure C-2d), assuming that areas with existing human activity will continue to expand, and designated them as barriers to elephant mobility. More details on these numbers can be found in section 4.5C.1.3 in 4.5Appendix C.

First, we used three primary habitat clusters as the focal node polygons in Circuitscape with pairwise modeling mode to imitate the connectivity between them. The environmental suitability acquired from multi-scale modeling (refer to section 3.3.3) was implemented as conductance in Circuitscape. The cumulative current map was rescaled to [0, 1] to indicate the long-distance landscape connectivity. Higher current density implies a greater likelihood of being utilized as a corridor, and therefore, being more critical in preserving landscape connectivity. We identified and trimmed the potential corridors between three habitat clusters by detecting the highest-current paths.

Next, we chose several protected areas (PAs) from the World Database on Protected Areas (WDPA) including Game Reserve, National Park, Game controlled area, Conservation Area, Open area, and Wildlife Management Area, as potential habitat units for elephants to simulate the connectivity within each habitat cluster. As certain PAs, including Open area, were subject to human-driven disturbances, we did not employ them as focal node polygons. Alternatively, we generated a random 5 km radius circle in each PA to represent a mini-habitat of elephants (Douglas-Hamilton et al., 2005; Shadrack et al., 2017), and used all circles as focal

nodes to run Circuitscape with pairwise modeling mode. We ran the simulation for 50 iterations, aggregated the cumulative current maps, and rescaled to [0, 1] as the short-range landscape connectivity within each primary habitat cluster. Employing this method, we were able to efficiently assess the effect of human disturbances on elephant population communication at a pixel level, even within PAs.

3.4 Results

3.4.1 Drivers of distributions and niches of African savanna elephants

At regional scales (2.5, 5, and 10 arc-minutes), settlement density, river density, and cropland edge density were found to be the most influential predictors for African savanna elephant distribution (Figure 3-3). The mean contiguity index ($CONTIG_{MN}$) of open habitat, which assesses the spatial connectedness of cells in patches (LaGro, 1991), was the fifth most important feature and the most significant landscape metric. These findings suggest that human modifications have a significant impact on elephant distribution in Tanzania at regional scales. Roads were found to have a slightly lower impact on elephant distribution than settlements and farms, possibly because roads do not necessarily have other human activities concentrated along them and can be shared by wildlife, particularly large mammals, under safe conditions. Among other natural resources, water availability is a directly decisive factor for elephant survival in savanna landscapes. For elephants, vegetation availability (NDVI) during the dry season is significantly less important than during the wet season and seasonal variability. One potential explanation is that during the dry season, elephants tend to prefer landscapes that are less

variable and remain green throughout the year, resulting in high NDVI during the wet season as well (Loarie et al., 2009). Moreover, their distribution in the dry season is limited to regions close to permanent water sources, which means that many well-wooded areas may be out of reach. It is also important to note that numerous trees in savanna woody areas shed their leaves during the dry season. As a result, these areas typically only exhibit higher NDVI than grasslands in the early dry season when grasses have dried, but leaves are still present, and their actual productivity benefits are not higher than those of grasslands. Climatic variables have relatively low influences on elephant distribution in savanna, as observed at the scales analyzed (Figure 3-4), mainly because elephants are adaptable to broad climate variability. Nevertheless, they still play some role, presumably due to their influence on other variables, such as forge and water availability (Dunkin et al., 2013).

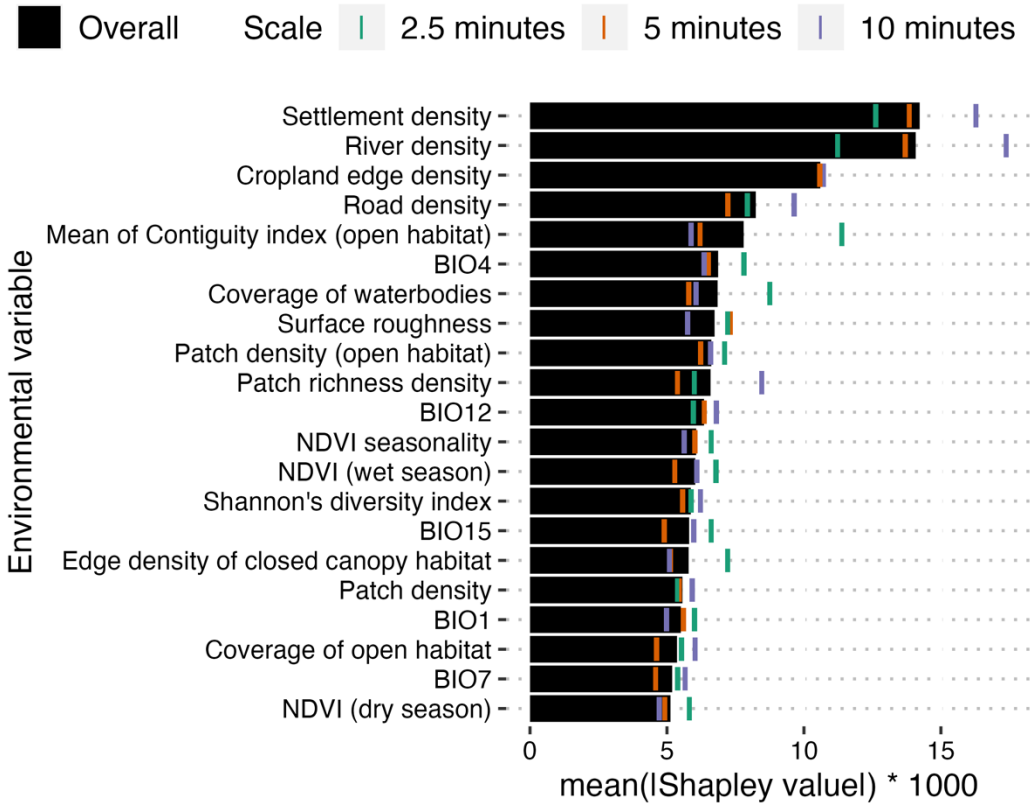

Figure 3-3. Variable importance at different coarse scales (10, 5, and 2.5 arc-minutes) for regional species distribution model (SDM). The mean of absolute Shapley values (x-axis) is used to indicate the importance of a variable in the SDM, with higher values indicating greater significance.

The univariate scaling analysis (section 3.3.3) identified an optimal window size of 3 km or 7 km for each variable at the landscape scale (Figure 3-4). This finding is consistent with the scaling effects observed at coarse scales (Figure 3-3). For example, the coverage of open habitat had higher predictive power at coarser scales, while the mean contiguity index of open habitat had higher predictive power at finer scales. Because the explanatory variables for landscape scale modeling were carefully selected based on multiple criteria, they all made significant contributions to the SDM (Figure 3-4). The coverage of open habitat with a window size of 7 km, the vector ruggedness measure (VRM) with a window size of 3 km, and the patch

density of open habitat with a window size of 3 km are the most important variables among those used. Vector ruggedness measure with a small window size is a dominant factor at the fine scale, demonstrating that topographic features play a significant role in the small-scale landscape selection of elephants. At landscape scales, resource factors such as the ratio of open habitat are more crucial than human activities like distance to settlements in determining the distribution of elephants, while at regional scales, the opposite appears to be true. Considering both regional and landscape scales, we can infer that human variables have a greater influence on elephant distribution at larger scales, while resource availability is the main driver of habitat selection at finer scales.

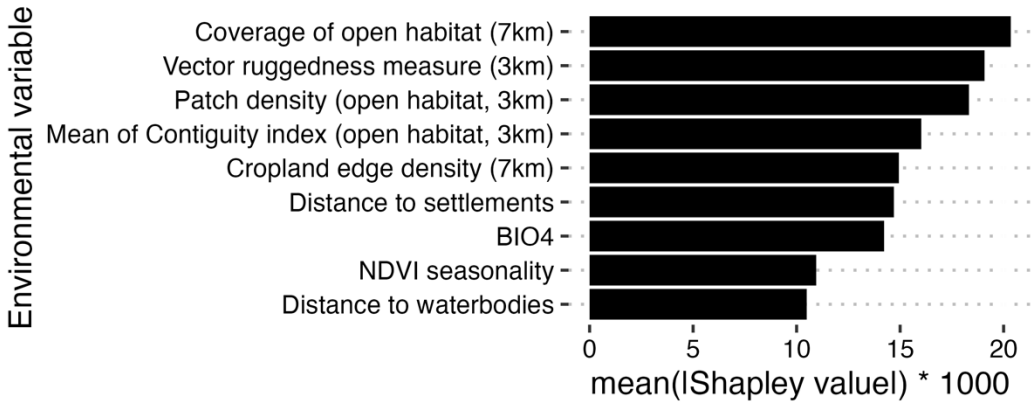

Figure 3-4. Variable importance at fine scale (1 km) for landscape species distribution model (SDM). #km is the window size of focal statistics. The mean of absolute Shapley values (x-axis) is used to indicate the importance of a variable in the SDM, with higher values indicating greater significance.

3.4.2 Environmental suitability of African savanna elephants

After the optimal scale analysis (Figure 3-2, and details are described in section 4.5C.2.1 in 4.5Appendix C), we chose 5 arc-minutes as the optimal scale for regional scale modeling. The regional-scale map (Figure 3-5a), generated and evaluated using pseudo-

occurrences with 10-fold cross-validation, showed an average AUC of 0.78 (SD = 0.02) and an average AUC_{ratio} of 0.74 (SD = 0.01) and an average F-measure of 0.76 (SD = 0.01). Using spatially thinned real occurrences (Figure 3-1) in 10-fold cross-validation (section 3.3.3), the landscape scale map (Figure 3-5b) had an average AUC of 0.8 (SD = 0.06), an average AUC_{ratio} of 0.82 (SD = 0.04), and an average F-measure of 0.77 (SD = 0.06). The integrated map (Figure 3-5c), evaluated using the same datasets in cross-validation of landscape scale modeling, had an average AUC of 0.82 (SD = 0.01), an average AUC_{ratio} of 0.94 (SD = 0.003), and an average F-measure of 0.77 (SD = 0.02). The results suggest that combining the regional scale map can enhance prediction accuracy and certainty at the landscape scale. Spatially, the map predicted by the fused model better represents overall suitability compared to individual models. Notably, even areas with high levels of agricultural activity, such as those in central Tanzania, are not entirely predicted as unsuitable for elephants, highlighting the possibility of severe human-elephant conflicts.

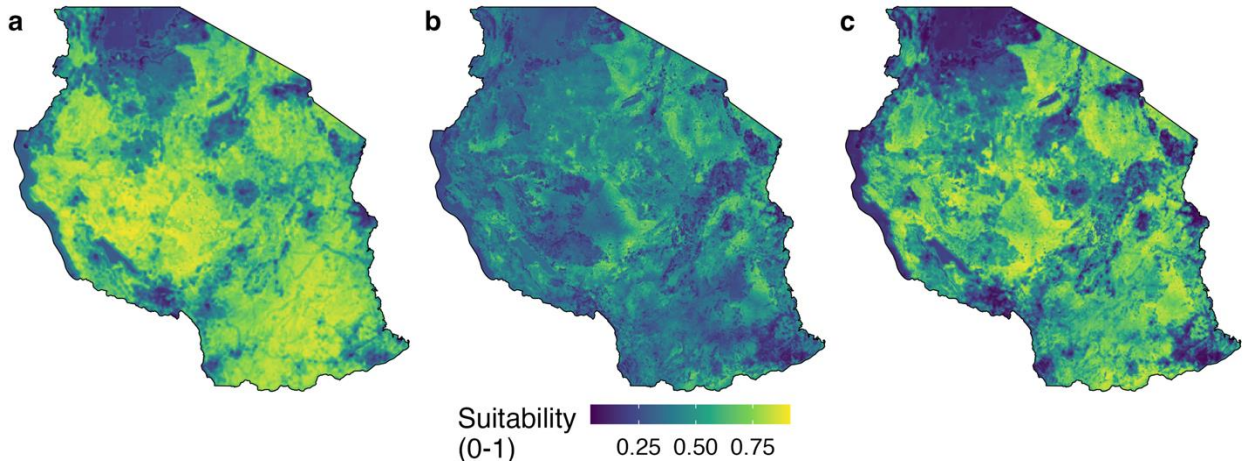

Figure 3-5. Environmental suitability map at the regional scale (5 arc-minutes, a), the landscape scale (0.5 arc-minutes, b), and Bayes fusion (c)

3.4.3 Landscape connectivity and potential corridors

3.4.3.1 Connectivity between habitat clusters

The combined result of a pairwise analysis conducted by Circuitscape, which evaluated the landscape connectivity between each primary habitat cluster to the other two under scenarios A and B, is displayed in Figure 3-6. Although currently facing the impacts of intensive human activities (Figure 3-6a and section 3.3.4), it has been predicted that the three primary habitat clusters of elephants will maintain some level of connectivity with each other, and five corridors have been identified as examples of such connectivity (Figure 3-6a). The area with the highest concentration of intensive human activities is North Tanzania, specifically across Lake Victoria and between habitat cluster No.1 and No.3 (Fig. 6a), which creates a complete blockage of connectivity in that region. Settlements and farmlands along the Tarangire National Park border, extending southeast to the Talamai Open Area (Figure 3-6a & I), are preventing elephants from leaving habitat cluster No.1 in this area. This could be the reason for the closure of corridors between Tarangire and Ruaha (Caro et al., 2009). On a positive note, it is predicted that a corridor (Figure 3-6a & 3-6III) between habitat cluster No.1 and No.3 may still be open. This corridor follows a path through Lake Kitangiri and its downstream wetlands (Figure 3-6III), and provides elephants with multiple food sources. However, this potential corridor is at a high risk of disappearing if agricultural expansion continues without proper ecological management (Figure 3-6b & 3-6III). At present, habitat clusters No.1 and No.2 are well connected (Figure 3-6a & 3-6IV). However, as agriculture continues to encroach upon critical areas suitable for elephant passage, landscape connectivity will be reshaped, and longer corridors will be required (Figure 3-6a & 3-6b). Multiple potential corridors have been identified for elephants to travel

between habitat cluster No.2 and No.3 (Figure 3-6a). These paths are relatively short but pass through undulating terrain. The mountain valleys along these paths are already being cultivated, which could severely disturb the connections between habitat cluster No.2 and No.3. Additionally, elephants in habitat cluster No.2 (Lohay et al., 2020) may find it uncomfortable to use these paths due to various disturbances surrounding the corridor entrances, including farms and settlements located in the valley between Uluguru Mountains, Malundwe Mountain, Udzungwa Mountain, Mbarika Mountains, and Mahenge Mountains (Figure 3-6II).

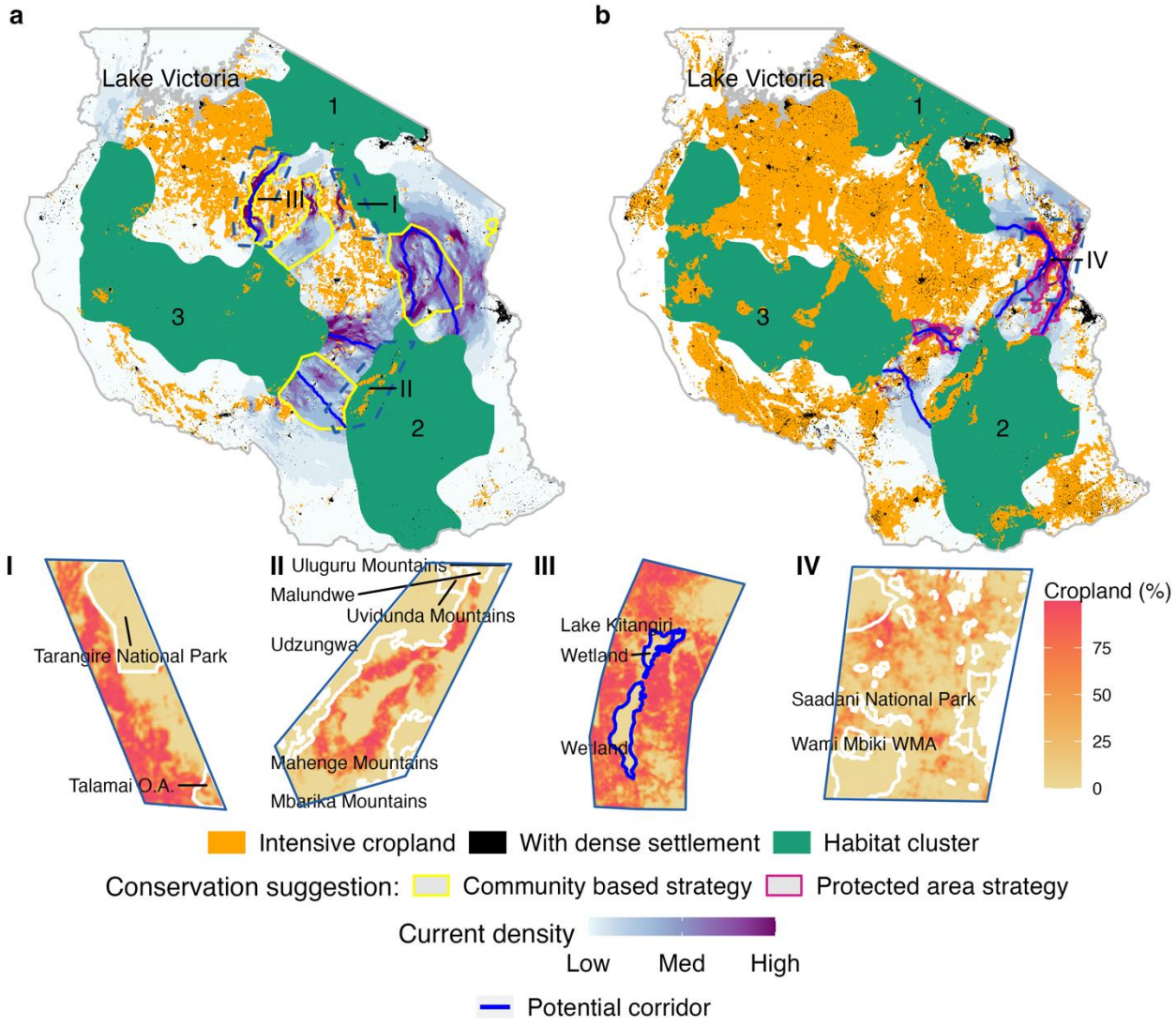

Figure 3-6. Landscape connectivity between primary habitat clusters (1, 2, & 3) under scenarios A (a) and B (b). In scenario A, pixels with >80% cropland coverage and containing >400 settlements are considered as barriers for elephant movement. In scenario B, pixels with >20% cropland coverage and containing >200 settlements are considered as potential barriers for elephant movement assuming current human disturbances will keep expanding over these areas. Areas I, II, III, and IV are highlighted areas.

3.4.3.2 Connectivity within habitat clusters

In a similar manner, we conducted a pairwise connectivity analysis between each PA (grey polygons in Figure 3-7) within each primary habitat cluster (No. 1-3 in Figure 3-6). The

resulting cumulative current maps were presented in Figure C-9 in 4.5Appendix C. To assess how potential agricultural and settlement expansion can spatially change landscape connectivity, we divided the cumulative current density predicted under scenario B by the cumulative current density predicted under scenario A. Figure 3-7 displays the results, which indicate that if human disturbances continue to expand, certain areas will lose their function as elephant population connectors, while other areas will become more important than before in maintaining elephant population connectivity.

Habitat cluster No.1 comprises Serengeti National Park, Ngorongoro Conservation Area, Lake Manyara National Park, Tarangire National Park, and several other PAs (Figure 3-7-1), but experiences varying levels of human disturbance in different regions. The areas between Tarangire and the open area in the north will remain conducive for movement, as long as there is no new agricultural expansion in this region. At present, elephants from Lake Manyara National Park and Tarangire National Park are able to move north and enter Ngorongoro Conservation Area (Figure 3-7-1) via Upper Kitete Corridor (Figure 3-7a, Lohay et al., 2022). However, regions leading up to Upper Kitete Corridor are being damaged rapidly due to permanent farms and houses. These regions include the contiguous farms between wards Mbuyuni and Madukani in the north of Tarangire (Figure 3-7c) and areas north of Lake Manyara, which also hinder communication between Lake Manyara and Tarangire National Park. There is another potential route for elephants to travel between Lake Manyara National Park and Ngorongoro Conservation Area, which involves passing through Marang Forest Reserve and the northeast lakeshore of Lake Eyasi, as shown in Figure 3-7a (Lohay et al., 2020). However, this corridor is currently largely obstructed due to the prevalence of intensive farming in the area.

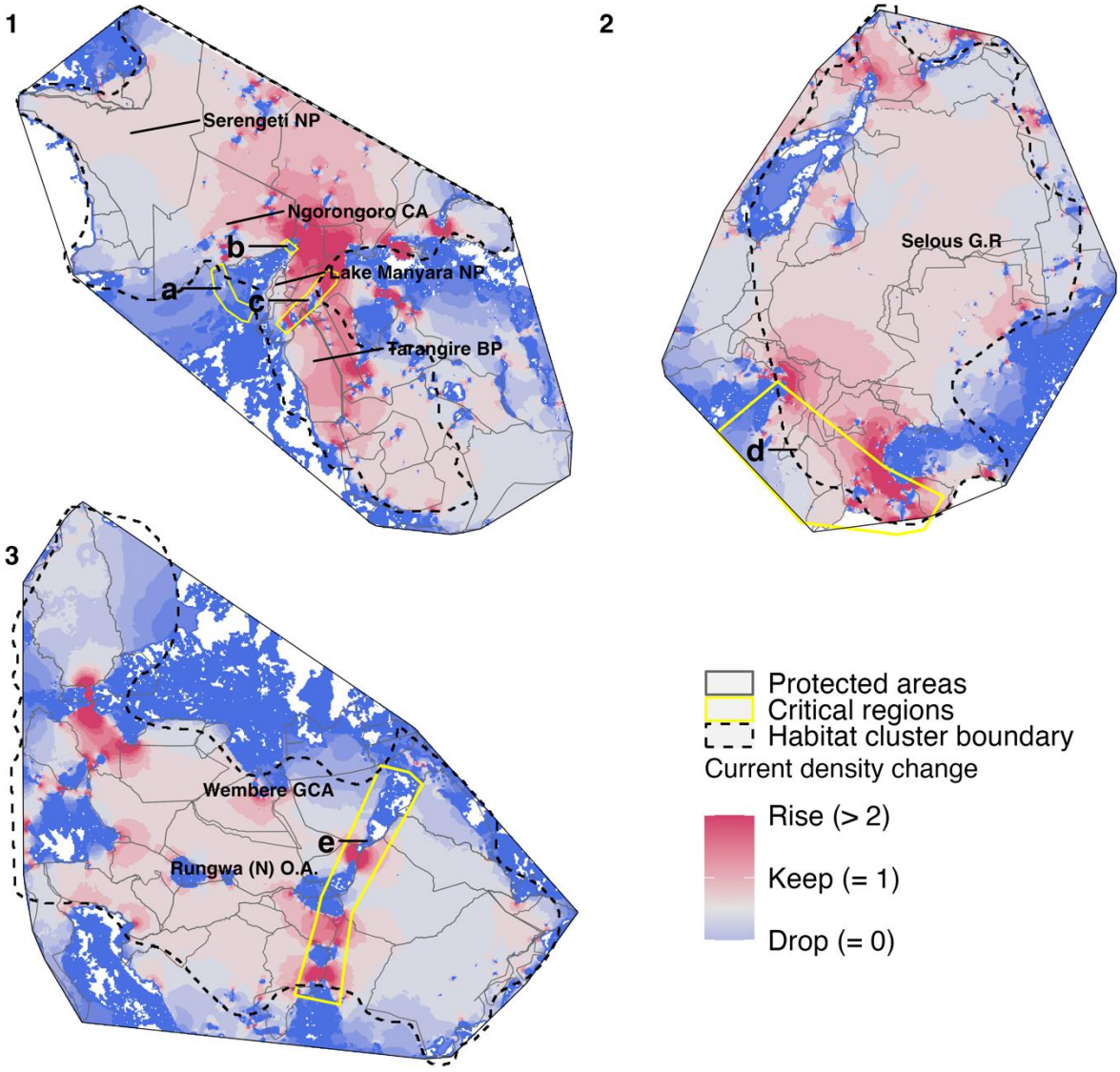

Figure 3-7. Difference in predicted landscape connectivity within each primary habitat cluster (No. 1-3 in Figure 3-6) under scenario A and B (section 3.3.4) and the detected critical regions of African elephant conservation in Tanzania. The current density changes were calculated by dividing the current density predicted under scenario B by the current density predicted under scenario A (Figure C-9 in 4.5Appendix C). The bluish color indicates areas with decreasing usage by elephants with the expansion of human activities, and the coral color indicates areas with increasing usage by elephants for movement, and thus become more critical for landscape connectivity.

Habitat cluster No.2 (Figure 3-7-2) encompasses a vast area and supports a significant population of elephants owing to the abundance of food and water resources. The elephant

populations within this habitat cluster experience relatively low levels of human disturbances, with the exception of the mountain valley located in the northwest, which is also indicated in Figure 3-6II. Currently, habitat cluster No.3 (Figure 3-7-3) is a vast and suitable habitat with good short-distance intra-connectivity for African elephants. However, if farming and housing continue to grow along the eastern border of Wembere Game Control Area and Rungwa Open Area (Figure 3-7e and Figure C-9), they may divide this habitat cluster into two parts, separating the elephants living in and around Ruaha National Park.

3.5 Discussion

3.5.1 Long-distance connectivity between primary habitat ecosystems

Despite the significant impact of human activities on the distribution and population connectivity of elephants in Tanzania (Figure 3-3 and Figure 3-4), their habitats still maintain varying degrees of connectivity. This offers promising options for implementing various strategies to preserve elephants' habitats and habitat connectivity in the face of rapid habitat fragmentation and loss (Jones et al., 2012). Since 2010, Tanzania gained notable increases in protected area connectivity (Saura et al., 2019). The landscape connectivity simulated in this study reinforces the findings of many other studies (Bond et al., 2017; Bukombe et al., 2022; Jones et al., 2012; Komba et al., 2021; Lohay et al., 2020; Martin et al., 2019; Riggio et al., 2018, 2022; Riggio & Caro, 2017), and offers spatially specific recommendations for further conservation management in Tanzania.

The simulation of long-distance landscape connectivity aided in identifying potential areas for various conservation practices to preserve habitat connectivity (Conservation suggestion in Figure 3-6). For regions already engaged in agricultural activities (e.g. targeted areas in Figure 3-6), implementing community-based conservation strategies, which have a long-standing history in Tanzania and have been effective in minimizing human-wildlife conflicts (Goldman, 2003; Salerno et al., 2016), is recommended. The Lake Kitangiri area (Fig. 6III) , for example, is particularly suitable because it is critical to link elephant populations in northern and southwestern habitats, and now is highly cultivated. For areas that are yet to be cultivated or settled, establishing new protected areas (such as the regions depicted in Figure 3-6IV) can safeguard critical hotspots for landscape connectivity and act as resourceful spots for long-distance migration. It is important to acknowledge that the remaining un-farmed habitat areas may already be utilized by people, especially pastoralists who use them for grazing their livestock (e.g. the open area in the middle of habitat cluster No. 1, Figure 3-7-1). Therefore, community-led approaches may be the most effective way to maintain these areas as grazing lands.

The Selous ecosystem is one of the biggest habitat and hosts a large population of elephants due to its abundant food and water resources. This may be one of the reasons why elephants in this region have less communication with other remote habitat clusters. There are clear connections between Selous and the Northern ecosystems (habitat cluster No. 2 in Fig. 6) through Wami-Mbiki (Figure 3-6IV) (Bukombe et al., 2022; Riggio et al., 2018).

There are evident existing connections from Selous through Wami-Mbiki (Figure 3-6IV) to link the Northern ecosystems (habitat cluster No. 2 in Figure 3-6), but the expansion of agriculture, illegal logging, and poaching pose significant threats to the essential connection

points along this path (Bukombe et al., 2022; Riggio et al., 2018). Wami-Mbiki Wildlife Management Area and its surrounding buffer zones can help maintain the necessary resource hub for elephant migration (Riggio et al., 2018), but additional protected areas are still required, such as areas as shown in Figure 3-6b. Elephant populations in the Selous ecosystem have been observed to move back and forth from Niassa Reserve in Mozambique (Lohay et al., 2020), making it crucial to keep the southern border of the Selous ecosystem open (Figure 3-7d).

3.5.2 Small-range connectivity priorities and conservation recommendations

As habitats become increasingly fragmented, seasonal connections between nearby patches of habitat have become particularly crucial for elephants to locate food. Our analysis of variables suggests that elephants tend to prefer areas with higher greenness (measured by higher NDVI). However, their distribution is limited to landscapes with access to permanent water sources during the dry season (as described in section 3.4.1). This finding has important implications for conservation efforts. Maintaining the availability of permanent water sources, such as by creating networks of artificial waterholes (Chamaillé-Jammes et al., 2007), between and near fragmented woody areas can enable elephants to access browsing habitats that were previously out of reach during the dry season. This can help sustain elephant populations and reduce browsing pressure on these woody habitats.

Human activities have heavily impacted the elephant habitats in the north of Tanzania, including the Tarangire-Manyara, Serengeti, and Ngorongoro ecosystems (habitat cluster No. 1 in Figure 3-6). The connection between Tarangire, Manyara, and nearby ecosystems is at high

risk of disappearing due to factors such as rapid agricultural and settlement expansion (Figure 3-7abc). Although the Upper Kitete Corridor remains open and can facilitate elephant migration from Tarangire and Manyara to the north such as Ngorongoro (Bond et al., 2017; Riggio et al., 2022; Riggio & Caro, 2017), the loss of corridors between Manyara and Tarangire in this area has been a concern for over 30 years (Mwalyosi, 1991). Thus, it is crucial to implement community-based conservation strategies in these areas to preserve elephant habitat connectivity and mitigate conflicts with local communities. The Marang Forest Reserve and northeast lakeshore of Lake Eyasi provide another potential corridor between Manyara and Ngorongoro (Figure 3-7a), which has been disrupted but can be restored through community-based farming practices in the surrounding areas.

Ruaha and the ecosystems surrounding it (habitat cluster No. 3 in Figure 3-6) currently offer stable elephant habitat. However, the fast-growing settlements and farmlands (Komba et al., 2021) increase the isolation of elephant population in Ruaha and Rungwa National Park (Figure 3-7e). This situation has been recognized and new protected areas have been established to maintain connectivity (Saura et al., 2019).

3.5.3 Conclusions and limitations

This paper presents a multi-scale modeling approach that utilizes environmental variables at different spatial scales and integrates multiple data sources that are different in design and accuracy (e.g. the expert range map and presence-only occurrences) to estimate species' environmental suitability. This approach is especially beneficial when modeling the distribution of species that have limited field observations. We also used Circuitscape to simulate

landscape connectivity at both broad and small scales within and between elephant habitats in Tanzania. The outcomes of this study provide spatially-explicit information that have important implications for conservation efforts.

However, there are several limitations to our approach. Firstly, elephants exhibit different habitat preferences during the dry and wet seasons, and their distribution is influenced by various factors. As a result, the corridors used by elephants may vary seasonally. Conducting separate simulations for wet and dry seasons could provide more accurate conservation insights, but would require extensive ground observations. Secondly, our model assumed that areas with existing human activities will continue to expand, and did not account for the complex nature of agricultural development in Tanzania, which involves intensification, expansion, and abandonment. Incorporating these factors into the model would require a massive amount of data, making it impractical or even impossible to implement. Recent advances in Earth Observation and Deep Learning may offer potential solutions in the future.

Data availability

All datasets and scripts are available via Open Science Framework (OSF): https://osf.io/5fwh4/ and GitHub: https://github.com/LLeiSong/eleDistribution.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was supported by the Future Investigators in NASA Earth and Space Science and Technology (FINESST) program (award number: 80NSSC20K1640).

References

Adriaensen, F., Chardon, J., De Blust, G., Swinnen, E., Villalba, S., Gulinck, H., & Matthysen, E. (2003). The application of 'least-cost'modelling as a functional landscape model. *Landscape and Urban Planning*, *64*(4), 233–247.

Alhajeri, B. H., & Fourcade, Y. (2019). High correlation between species-level environmental data estimates extracted from IUCN expert range maps and from GBIF occurrence data. *Journal of Biogeography*, jbi.13619. https://doi.org/10.1111/jbi.13619

Berger, J. (2004). The Last Mile: How to Sustain Long-Distance Migration in Mammals. *Conservation Biology*, *18*(2), 320–331. https://doi.org/10.1111/j.1523-1739.2004.00548.x

Berrar, D. (2018). Bayes' theorem and naive Bayes classifier. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, *403*.

Bond, M. L., Bradley, C. M., Kiffner, C., Morrison, T. A., & Lee, D. E. (2017). A multi-method approach to delineate and validate migratory corridors. *Landscape Ecology*, *32*(8), 1705–1721. https://doi.org/10.1007/s10980-017-0537-4

Bukombe, J., Marealle, W., Kimaro, J., Kija, H., Kavana, P., Kakengi, V., Nindi, J., Keyyu, J., Ntalwila, J., Kilimba, N., Bwenge, F., Nkwabi, A., Lowassa, A., Sanare, J., Mwita, M., Leweri, C., Kohi, E., Mangewa, L., Juma, R., … Lobora, A. (2022). Viability assessment of the Wami-Mbiki Game Reserve to Nyerere National Park wildlife corridor in southern Tanzania. *Global Ecology and Conservation*, *39*, e02259. https://doi.org/10.1016/j.gecco.2022.e02259

Caro, T., Jones, T., & Davenport, T. R. B. (2009). Realities of documenting wildlife corridors in tropical countries. *Biological Conservation*, *142*(11), 2807–2811. https://doi.org/10.1016/j.biocon.2009.06.011

Chamberlain, S., Barve, V., Mcglinn, D., Oldoni, D., Desmet, P., Geffert, L., & Ram, K. (2022). *rgbif: Interface to the Global Biodiversity Information Facility API*. https://CRAN.R-project.org/package=rgbif

Chamberlain, S., & Boettiger, C. (2017). R Python, and Ruby clients for GBIF species occurrence data. *PeerJ PrePrints*. https://doi.org/10.7287/peerj.preprints.3304v1

Cortes, D. (2021a). *isotree: Isolation-Based Outlier Detection*. https://CRAN.R-project.org/package=isotree

Cortes, D. (2021b). Revisiting randomized choices in isolation forests. *ArXiv:2110.13402 [Cs, Stat]*. http://arxiv.org/abs/2110.13402

de Boer, W. F., van Langevelde, F., Prins, H. H. T., de Ruiter, P. C., Blanc, J., Vis, M. J. P., Gaston, K. J., & Hamilton, I. D. (2013). Understanding spatial differences in African elephant densities and occurrence, a continent-wide analysis. *Biological Conservation*, *159*, 468–476. https://doi.org/10.1016/j.biocon.2012.10.015

Dejene, S. W., Mpakairi, K. S., Kanagaraj, R., Wato, Y. A., & Mengistu, S. (2021). Modelling continental range shift of the African elephant (*Loxodonta africana*) under a changing climate and land cover: Implications for future conservation of the species. *African Zoology*, *56*(1), 25–34. https://doi.org/10.1080/15627020.2020.1846617

Douglas-Hamilton, I. (1987). African elephants: Population trends and their causes. *Oryx*, *21*(1), 11–24.

Douglas-Hamilton, I., Krink, T., & Vollrath, F. (2005). Movements and corridors of African elephants in relation to protected areas. *Naturwissenschaften*, *92*(4), 158–163. https://doi.org/10.1007/s00114-004-0606-9

Drake, J. M., & Richards, R. L. (2018). Estimating environmental suitability. *Ecosphere*, *9*(9), e02373. https://doi.org/10.1002/ecs2.2373

Dunkin, R. C., Wilson, D., Way, N., Johnson, K., & Williams, T. M. (2013). Climate influences thermal balance and water use in African and Asian elephants: Physiology can predict drivers of elephant distribution. *Journal of Experimental Biology*, *216*(15), 2939–2952.

Epps, C. W., Wasser, S. K., Keim, J. L., Mutayoba, B. M., & Brashares, J. S. (2013). Quantifying past and present connectivity illuminates a rapidly changing landscape for the African elephant. *Molecular Ecology*, *22*(6), 1574–1588. https://doi.org/10.1111/mec.12198

Fourcade, Y. (2016). Comparing species distributions modelled from occurrence data and from expert-based range maps. Implication for predicting range shifts with climate change. *Ecological Informatics*, *36*, 8–14. https://doi.org/10.1016/j.ecoinf.2016.09.002

Fryxell, J., & Sinclair, A. (1988). Causes and consequences of migration by large herbivores. *Trends in Ecology & Evolution*, *3*(9), 237–241.

Gara, T. W., Wang, T., Dube, T., Ngene, S. M., & Mpakairi, K. S. (2021). African elephant (*Loxodonta africana*) select less fragmented landscapes to connect core habitats in human-dominated landscapes. *African Journal of Ecology*, *59*(2), 370–377. https://doi.org/10.1111/aje.12839

Gobush, K. S., Edwards, C. T. T., Balfour, D., Wittemyer, G., Maisels, F., & Taylor, R. D. (2022). *Loxodonta africana (amended version of 2021 assessment)*. The IUCN Red List of Threatened Species 2022: E.T181008073A223031019. https://dx.doi.org/10.2305/IUCN.UK.2022-2.RLTS.T181008073A223031019.en

Goheen, J. R., Palmer, T. M., Keesing, F., Riginos, C., & Young, T. P. (2010). Large herbivores facilitate savanna tree establishment via diverse and indirect pathways. *Journal of Animal Ecology*, *79*(2), 372–382. https://doi.org/10.1111/j.1365-2656.2009.01644.x

Green, S. E., Davidson, Z., Kaaria, T., & Doncaster, C. P. (2018). Do wildlife corridors link or extend habitat? Insights from elephant use of a Kenyan wildlife corridor. *African Journal of Ecology*, *56*(4), 860–871. https://doi.org/10.1111/aje.12541

Guha, S., Mishra, N., Roy, G., & Schrijvers, O. (2016). Robust random cut forest based anomaly detection on streams. *International Conference on Machine Learning*, 2712–2721.

Hall, K. R., Anantharaman, R., Landau, V. A., Clark, M., Dickson, B. G., Jones, A., Platt, J., Edelman, A., & Shah, V. B. (2021). Circuitscape in Julia: Empowering Dynamic Approaches to Connectivity Assessment. *Land*, *10*(3), 301. https://doi.org/10.3390/land10030301

Hanley, J. A. & others. (1989). Receiver operating characteristic (ROC) methodology: The state of the art. *Crit Rev Diagn Imaging*, *29*(3), 307–335.

Hauke, J., & Kossowski, T. (2011). Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones Geographicae*, *30*(2), 87.

Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C., & Guisan, A. (2006). Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, *199*(2), 142–152. https://doi.org/10.1016/j.ecolmodel.2006.05.017

Jones, T., Bamford, A. J., Ferrol-Schulte, D., Hieronimo, P., McWilliam, N., & Rovero, F. (2012). Vanishing Wildlife Corridors and Options for Restoration: A Case Study from Tanzania. *Tropical Conservation Science*, *5*(4), 463–474. https://doi.org/10.1177/194008291200500405

Kiffner, C. (2022). *Tarangire: Human-Wildlife Coexistence in a Fragmented Ecosystem*. Springer Nature.

Kohi, E. M., de Boer, W. F., Peel, M. J. S., Slotow, R., van der Waal, C., Heitkönig, I. M. A., Skidmore, A., & Prins, H. H. T. (2011). African Elephants Loxodonta africana Amplify Browse Heterogeneity in African Savanna: Elephants Amplify Browse Heterogeneity. *Biotropica*, *43*(6), 711–721. https://doi.org/10.1111/j.1744-7429.2010.00724.x

Komba, A. W., Watanabe, T., Kaneko, M., & Chand, M. B. (2021). Monitoring of Vegetation Disturbance around Protected Areas in Central Tanzania Using Landsat Time-Series Data. *Remote Sensing*, *13*(9), 1800. https://doi.org/10.3390/rs13091800

LaGro, J. (1991). Assessing patch shape in landscape mosaics. *Photogrammetric Engineering and Remote Sensing*, *57*(3), 285–293.

Laursen, L., & Bekoff, M. (1978). Loxodonta africana. *Mammalian Species*, *92*, 1. https://doi.org/10.2307/3503889

Leroy, B., Delsol, R., Hugueny, B., Meynard, C. N., Barhoumi, C., Barbet-Massin, M., & Bellard, C. (2018). Without quality presence–absence data, discrimination metrics such as TSS can be misleading measures of model performance. *Journal of Biogeography*, *45*(9), 1994–2002.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. *2008 Eighth IEEE International Conference on Data Mining*, 413–422.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-Based Anomaly Detection. *ACM Transactions on Knowledge Discovery from Data*, *6*(1), 1–39. https://doi.org/10.1145/2133360.2133363

Lohay, G. G., Riggio, J., Lobora, A. L., Kissui, B. M., & Morrison, T. A. (2022). Wildlife movements and landscape connectivity in the Tarangire Ecosystem. In *Tarangire: Human-Wildlife Coexistence in a Fragmented Ecosystem* (pp. 255–276). Springer.

Lohay, G. G., Weathers, T. C., Estes, A. B., McGrath, B. C., & Cavener, D. R. (2020). Genetic connectivity and population structure of African savanna elephants (*Loxodonta africana*) in Tanzania. *Ecology and Evolution*, *10*(20), 11069–11089. https://doi.org/10.1002/ece3.6728

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777.

Macdonald, D. W., Boitani, L., Dinerstein, E., Fritz, H., & Wrangham, R. (2013). Conserving large mammals: Are they a special case? *Key Topics in Conservation Biology 2*, 277–312.

Martin, E. H., Jensen, R. R., Hardin, P. J., Kisingo, A. W., Shoo, R. A., & Eustace, A. (2019). Assessing changes in Tanzania's Kwakuchinja Wildlife Corridor using multitemporal satellite imagery and open source tools. *Applied Geography*, *110*, 102051. https://doi.org/10.1016/j.apgeog.2019.102051

Massey Jr, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, *46*(253), 68–78.

McRae, B. H., Dickson, B. G., Keitt, T. H., & Shah, V. B. (2008). Using circuit theory to model connectivity in ecology, evolution, and conservation. *Ecology*, *89*(10), 2712–2724. https://doi.org/10.1890/07-1861.1

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com. https://christophm.github.io/interpretable-ml-book/

Mpakairi, K. S., Ndaimani, H., Kuvawoga, P. T., & Madiri, H. T. (2019). Human settlement drives African elephant (*Loxodonta africana*) movement in the Sebungwe Region, Zimbabwe. *African Journal of Ecology*, *57*(4), 531–538. https://doi.org/10.1111/aje.12639

Newmark, W. D. (2008). Isolation of African protected areas. *Frontiers in Ecology and the Environment*, *6*(6), 321–328.

Norway's International Climate and Forest Initiative (NICFI). (2020, May 18). NICFI. https://www.nicfi.no/

Ntukey, L. T., Munishi, L. K., Kohi, E., & Treydte, A. C. (2022). Land Use/Cover Change Reduces Elephant Habitat Suitability in the Wami Mbiki–Saadani Wildlife Corridor, Tanzania. *Land*, *11*(2), 307. https://doi.org/10.3390/land11020307

Peterson, A. T., Papeş, M., & Soberón, J. (2008). Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling*, *213*(1), 63–72. https://doi.org/10.1016/j.ecolmodel.2007.11.008

Pringle, R. M., Prior, K. M., Palmer, T. M., Young, T. P., & Goheen, J. R. (2016). Large herbivores promote habitat specialization and beta diversity of African savanna trees. *Ecology*, *97*(10), 2640–2657.

Purdon, A., Mole, M. A., Chase, M. J., & van Aarde, R. J. (2018). Partial migration in savanna elephant populations distributed across southern Africa. *Scientific Reports*, *8*(1), 11331. https://doi.org/10.1038/s41598-018-29724-9

Riggio, J., & Caro, T. (2017). Structural connectivity at a national scale: Wildlife corridors in Tanzania. *PLOS ONE*, *12*(11), e0187407. https://doi.org/10.1371/journal.pone.0187407

Riggio, J., Foreman, K., Freedman, E., Gottlieb, B., Hendler, D., Radomille, D., Rodriguez, R., Yamashita, T., Kioko, J., & Kiffner, C. (2022). Predicting wildlife corridors for multiple species in an East African ungulate community. *PLOS ONE*, *17*(4), e0265136. https://doi.org/10.1371/journal.pone.0265136

Riggio, J., Mbwilo, F., Van de Perre, F., & Caro, T. (2018). The forgotten link between northern and southern Tanzania. *African Journal of Ecology*, *56*(4), 1012–1016. https://doi.org/10.1111/aje.12533

Rotenberry, J. T., & Balasubramaniam, P. (2020). Connecting species' geographical distributions to environmental variables: Range maps versus observed points of occurrence. *Ecography*, *43*(6), 897–913. https://doi.org/10.1111/ecog.04871

Rowhani, P., Lobell, D. B., Linderman, M., & Ramankutty, N. (2011). Climate variability and crop production in Tanzania. *Agricultural and Forest Meteorology*, *151*(4), 449–460. https://doi.org/10.1016/j.agrformet.2010.12.002

Shadrack, N., Moses, M. O., Joseph, M., Shadrack, M., Steve, N., & James, I. (2017). Home range sizes and space use of African elephants (Loxodonta africana) in the Southern Kenya and Northern Tanzania borderland landscape. *International Journal of Biodiversity and Conservation*, *9*(1), 9–26. https://doi.org/10.5897/IJBC2016.1033

Shapley, L. S. (1953). A value for n-person games. In *Contributions to the Theory of Games* (2.28, pp. 307–317).

Shen, Y., Liu, M., Wang, D., Shen, X., & Li, S. (2021). Using an integrative mapping approach to identify the distribution range and conservation needs of a large threatened mammal, the Asiatic black bear, in China. *Global Ecology and Conservation*, *31*, e01831. https://doi.org/10.1016/j.gecco.2021.e01831

Song, L., Estes, A. B., & Estes, L. D. (2023). A super-ensemble approach to map land cover types with high resolution over data-sparse African savanna landscapes. *International Journal of Applied Earth Observation and Geoinformation*, *116*, 103152. https://doi.org/10.1016/j.jag.2022.103152

Sun, X., Long, Z., & Jia, J. (2021). A multi-scale Maxent approach to model habitat suitability for the giant pandas in the Qionglai mountain, China. *Global Ecology and Conservation*, *30*, e01766. https://doi.org/10.1016/j.gecco.2021.e01766

Szczys, P., Oswald, S. A., & Arnold, J. M. (2017). Conservation implications of long-distance migration routes: Regional metapopulation structure, asymmetrical dispersal, and population declines. *Biological Conservation*, *209*, 263–272. https://doi.org/10.1016/j.biocon.2017.02.012

Tanzania National Bureau of Statistics. (2021). *Tanzania in Figures 2021*. Tanzania National Bureau of Statistics. https://www.nbs.go.tz/index.php/en/tanzania-in-figures/784-tanzania-in-figures-2021

Thouless, C., Dublin, H. T., Blanc, J., Skinner, D., Daniel, T., Taylor, R., Maisels, F., Frederick, H., & Bouché, P. (2016). African elephant status report 2016. *Occasional Paper Series of the IUCN Species Survival Commission*, *60*.

Timm, B. C., McGarigal, K., Cushman, S. A., & Ganey, J. L. (2016). Multi-scale Mexican spotted owl (Strix occidentalis lucida) nest/roost habitat selection in Arizona and a comparison with single-scale modeling results. *Landscape Ecology*, *31*(6), 1209–1225. https://doi.org/10.1007/s10980-016-0371-0

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., & others. (2012). GMES Sentinel-1 mission. *Remote Sensing of Environment*, *120*, 9–24.

Tshipa, A., Valls-Fox, H., Fritz, H., Collins, K., Sebele, L., Mundy, P., & Chamaillé-Jammes, S. (2017). Partial migration links local surface-water management to large-scale elephant conservation in the world's largest transfrontier conservation area. *Biological Conservation*, *215*, 46–50. https://doi.org/10.1016/j.biocon.2017.09.003

Turner, M. G., & Gardner, R. H. (2015). *Landscape ecology in theory and practice: Pattern and process* (Second edition). Springer.

Vasudev, D., Goswami, V. R., Srinivas, N., Syiem, B. L. N., & Sarma, A. (2021). Identifying important connectivity areas for the wide-ranging Asian elephant across conservation landscapes of Northeast India. *Diversity and Distributions*, *27*(12), 2510–2526. https://doi.org/10.1111/ddi.13419

Vidal, M. M., Pires, M. M., & Guimarães, P. R. (2013). Large vertebrates as the missing components of seed-dispersal networks. *Biological Conservation*, *163*, 42–48. https://doi.org/10.1016/j.biocon.2013.03.025

Wall, J., Douglas-Hamilton, I., & Vollrath, F. (2006). Elephants avoid costly mountaineering. *Current Biology*, *16*(14), R527–R529. https://doi.org/10.1016/j.cub.2006.06.049

Wall, J., Wittemyer, G., Klinkenberg, B., LeMay, V., & Douglas-Hamilton, I. (2013). Characterizing properties and drivers of long distance movements by elephants (Loxodonta africana) in the Gourma, Mali. *Biological Conservation*, *157*, 60–68. https://doi.org/10.1016/j.biocon.2012.07.019

Williams, H. F., Bartholomew, D. C., Amakobe, B., & Githiru, M. (2018). Environmental factors affecting the distribution of African elephants in the Kasigau wildlife corridor, SE Kenya. *African Journal of Ecology*, *56*(2), 244–253. https://doi.org/10.1111/aje.12442

Wittemyer, G., Northrup, J. M., Blanc, J., Douglas-Hamilton, I., Omondi, P., & Burnham, K. P. (2014). Illegal killing for ivory drives global decline in African elephants. *Proceedings of the National Academy of Sciences*, *111*(36), 13117–13121. https://doi.org/10.1073/pnas.1403984111

Chapter 4

Cropland allocation to minimize agriculture-elephant conflict with consideration of biodiversity and carbon costs

4.1 Abstract

With an increasing population and economics, sub-Saharan Africa (SSA) is undergoing intensive agricultural expansion, resulting in habitat loss, degradation, and fragmentation. Human activities, such as agriculture, act as significant barriers to block habitat connectivity, particularly for large mammals, within these highly fragmented landscapes. This has led to a reduction in effective size of species populations, hindered gene flow between populations, and increased human-wildlife conflicts. There are efforts to try minimizing the impact of ongoing fragmentation on wildlife populations by undertaking land use modeling and tradeoff analyses, which enable human development needs and conservation objectives to be better balanced. While previous studies of tradeoffs between agriculture and biodiversity mainly rely on occupancy-based assessment, such as species richness, with less land available for conservation, it is becoming essential to consider landscape connectivity in such analyses. In this study, we applied species distribution modeling (SDM) and circuit theory to estimate the spatial importance of habitat connectivity for African elephants and incorporated this information into a model designed to allocate land for agricultural expansion. We also included biodiversity and

carbon cost in the model to satisfy other important conservation criteria. By setting different weights to different criteria, the model can target potential land for agricultural expansion and achieve quantitatively optimal trade-offs between competing demands. Our analysis suggests that closing crop yield gaps and maximizing the utilization of presently cultivated lands, such as expanding farm sizes, may be effective solutions for managing the growing food demand in Tanzania. The model results provide guidance for land use planning that helps achieve production targets while conserving ecosystems.

Keywords: Human-wildlife conflict, elephant conservation, biodiversity, carbon, agricultural expansion, agricultural intensification

4.2 Introduction

The twenty-first century presents a significant challenge of balancing food demands with the preservation of ecosystem services. Habitat loss poses a significant threat to wild species, as it reduces the size of the area that they can occupy and fragments their populations and ranges into small, isolated patches, which arguably accelerates the risk of extinction for these species (Fastré et al., 2020; Rands et al., 2010). The increasing human population and its corresponding per capita consumption have led to the prevalent expansion of agriculture, which has become the primary driver of biodiversity loss and destruction and degradation of natural ecosystems (Dudley & Alexander, 2017; Haddad et al., 2015). Agriculture-driven habitat loss currently represents the most severe threat to terrestrial vertebrates, with 20% of these species facing the possibility of extinction (Tilman et al., 2017). These threats are likely to persist, as projected increases in human populations are expected to drive additional land-use changes

(Winkler et al., 2021). Particularly, sub-Saharan Africa (SSA) is likely to be swept by a projected doubling of population by 2050 combined with rapidly growing economies (Estes et al., 2016; Tilman et al., 2017; D. R. Williams et al., 2020). The greatest elevated extinction risks are predicted to be experienced by mammals in SSA (Tilman et al., 2017) because large-body species are especially vulnerable to human-driven impacts (Cardillo et al., 2005; Ripple et al., 2015). East Africa, as a biodiversity hotspot in SSA with a rapidly growing population and expanding agriculture (Doggart et al., 2020; Tilman et al., 2017; D. R. Williams et al., 2020), requires urgent attention to address both ecological conservation and food security.

The extent to which human activities overlap with biodiversity is a critical determinant in assessing the impact of human activities on biodiversity. The required amount of land to meet a country's future crop demands will depend on the land's yields. Closing yield gaps could substantially lower the need for future land clearance (Mueller et al., 2012; Tilman et al., 2017). More proactive and concerted conservation strategies, for example, shifts in agricultural practices, will also be essential to slow global biodiversity declines (Travers et al., 2019; D. R. Williams et al., 2020). Therefore, understanding the spatially explicit distribution of high-profit agricultural practices and their potential ecological consequences is crucial for minimizing the negative impact of agriculture on biodiversity and ecosystems while maximizing its productive benefits (D. R. Williams et al., 2020).

A substantial amount of work focuses on reconciliation between biodiversity conservation and agricultural development. Occupancy-based species measurements, such as species richness (Brancalion et al., 2019; Brooks et al., 2019), species rarity (Estes et al., 2016), ecosystem intactness (Mokany et al., 2020; Soto-Navarro et al., 2020), and extinction risk (Strassburg et al., 2018; B. A. Williams et al., 2020), are commonly employed in this body of

studies to evaluate ecological costs of land conversion and impacts on biodiversity changes. Nonetheless, despite the growing recognition of the importance of landscape connectivity in biodiversity conservation in the face of human-induced habitat degradation and fragmentation, current research has neglected to incorporate this aspect in their assessments (Newbold et al., 2015). By considering the flow of resources, organisms, and genetic material, landscape-scale approaches can provide a more comprehensive understanding of the interconnections between habitats and the potential ecological consequences of land-use change (Rands et al., 2010). The integration of species distribution modeling (SDM) and landscape connectivity analysis offers a solution to this problem. SDM links species occurrence data to environmental variables and creates species' spatial environmental suitability (Elith et al., 2006; Elith & Leathwick, 2009), which can serve as a base layer to analyze landscape connectivity. Various techniques, such as graph-based methods (Hall et al., 2021; Hofman et al., 2018), are employed in landscape connectivity analysis to map out habitat networks and evaluate habitat connectivity. The utilization of these techniques has garnered positive outcomes in urban planning (such as Morin et al., 2022; Tarabon et al., 2019) and in identifying core areas for habitat restoration (such as Clauzel & Godet, 2020; Préau et al., 2022). Thus, it is imperative that these methods be integrated into the tradeoff analysis between agriculture and biodiversity, particularly considering large-size terrestrial mammals (such as African elephants).

Here, we explicitly evaluated the landscape connectivity of African savanna elephants and integrated it into the reconciliation framework, along with measures of biodiversity and carbon cost, to optimize agricultural land allocation and minimize conflicts between agriculture and elephants. Our study aimed to balance the criteria of agriculture and conservation by incorporating the crucial aspect of landscape connectivity into the decision-making process.

Additionally, the spatial analysis model (Estes et al., 2016) enables the evaluation of varying levels of compromise between competing interests and still meet agricultural production targets. Our findings can suggest optimal areas for agricultural expansion with minimal ecological costs and achieve a sustainable land use solution. By considering both ecological and agricultural interests, our approach can help in making informed decisions for future land use planning.

4.3 Materials and methods

4.3.1 Study region and objectives

The United Republic of Tanzania (Figure 4-1) is the largest country located in East Africa, covering an area of approximately 947,300 km^2 (Tanzania National Bureau of Statistics, 2021), and has a diverse geography and climate (Luhunga et al., 2018). It is known for various species and high biodiversity, but it also faces increasing challenges in balancing agricultural expansion and human-wildlife conflicts. Elephants, as one of the largest and widely distributed species in Tanzania, are largely threatened by habitat loss, fragmentation, and degradation, and often in conflict with local communities over crop damage and destruction of property (Shaffer et al., 2019). The challenges of balancing agricultural development and elephant conservation are complex and require a holistic approach that involves the engagement of local communities, government agencies, and other stakeholders (Dickman, 2010; Hoare, 2015; Shaffer et al., 2019). A proper balance is necessary to ensure both sustainable agricultural development and the conservation of elephants and other wildlife species in Tanzania.

Consequently, this study aims to optimize land allocation to satisfy the production targets for staple food crops, while minimizing the impact on elephant conservation as quantified

by habitat connectivity and considering biodiversity and carbon costs. To accomplish this, we evaluated the agricultural production and potential, elephant migration conservation, biodiversity cost, and carbon cost for land conversion in each 1 km planning unit across the entire country. Then, using a spatial tradeoff analysis model (Estes et al., 2016), we allocated farming areas with the highest potential agricultural production and the lowest environmental costs for new croplands.

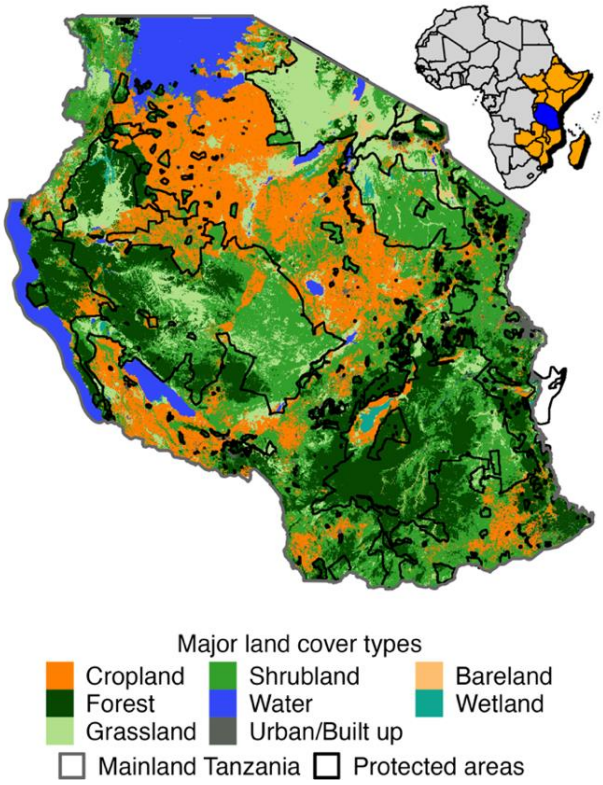

Figure 4-1. Spatial location and geographic characteristics of Tanzania.

Our analysis evaluated the overall agricultural production with focus on multiple staple crops, maize, rice, cassava and pulses, due to their current and future importance to Tanzania's agriculture and the expected growth in their cultivation (Rowhani et al., 2011; URT, 2021). We used 5×10^7 tonnes (roughly 4-folds of current production) as the future demand of selected

crops for the land allocation. Details on model structure and inputs are described in the following sections.

4.3.2 Agricultural expansion and intensification

We conducted a spatial analysis to determine the production gain of current cropland from agricultural intensification and non-cropland from expansion in order to meet future food demands. To determine the current yield of selected crops, we used the GAEZ+ 2015 annual crop dataset with a 5-minute resolution (Grogan et al., 2022). The rainfed crop yield maps were downscaled to 1 km resolution using a Random Forest (RF) model and relevant variables, including vegetation indices, meteorological variables, and soil predictors. The vegetation indices were the mean Normalized Difference Vegetation Index (NDVI), Gross Primary Product (GPP), and Leaf Area Index (LAI) (Lobell et al., 2015) while the meteorological variables included mean evapotranspiration, precipitation, temperature, maximum temperature, minimum temperature, and water vapor pressure (Fick & Hijmans, 2017) during the growing season (November to May). NDVI, GPP, LAI, and evapotranspiration were driven from MODIS (Myneni et al., 2015; Running et al., 2015, 2021) in Google Earth Engine. The soil predictors consisted of the mean values of bulk density, cation exchange capacity, coarse fragments volumetric, clay content, nitrogen, pH in H_2O, sand content, silt content, and soil organic carbon content across all layers (Poggio et al., 2021).

Similarly, we queried the simulated near future (2011-2040) agro-ecological water-limited attainable yield maps at 5-minute resolution from GAEZ v4 data with all climate data sources and calculated the average as the attainable yields (Fischer et al., 2021). To match with

current yield maps, we calculated the maximum yield that can be attained by chickpea, cow pea, pearl millet, Phaseolus bean, and pigeon pea as the attainable yield of pulses (Grogan et al., 2022), and the average attainable yield for dryland and wetland rice as the attainable yield of rice. These maps were then downscaled to a 1 km resolution with a Random Forest model and relevant meteorological and soil variables, which are the same soil predictors as in the current yield downscaling (Poggio et al., 2021). Meteorological variables such as the growing-season mean of future (2021-2040) precipitation, temperature, maximum temperature, and minimum temperature were obtained from the Worldclim website (https://www.worldclim.org) and calculated across all global climate models under scenario ssp245 (Fick & Hijmans, 2017). More details and maps can be found in section 4.5D.1 in 4.5Appendix D.

The overall current yield (Yc) and attainable yield (Ya) for each planning unit i are calculated using the downscaled yield and attainable yield maps as follows:

$$Yc_i = \frac{\sum_{c \in crops} Yc_i(c)}{n} \tag{4-1}$$

$$Ya_i = \frac{\sum_{c \in crops} Ya_i(c)}{n} \tag{4-2}$$

where crops include cassava, maize, rice and pulses and n is the number of crops ($n = 4$).

Consequently, the production gain from agricultural intensification (Pig) in each planning unit i that encompasses cultivated areas is calculated as follows:

$$Pig_i = (Ya_i - Yc_i) \times Ac_i \tag{4-3}$$

where Ac_i is the area of current cropland within each planning unit. Yig_i can be theoretically negative in areas that are predicted to become less productive in the future due to climate change.

The production gain from agricultural expansion (Peg) is calculated differently for planning unit i with (Uc) or without (Un) current cropland:

$$Peg_i = \begin{cases} Ya_i \times A_m, i \in Un \\ Ya_i \times (\max(0, A_m - Ac_i)), i \in Uc \end{cases} \quad (4\text{-}4)$$

where A_m is the allowed maximum area to convert to cropland in each planning unit.

The benefit of agricultural development is represented by the sum of production gains from intensification and expansion, the layer of attainable yield, as well as a map of travel time to cities (Weiss et al., 2018), which serve as decision-making factors to represent production benefits in the spatial analysis. The distribution of attainable production gain is used as the baseline to determine the total amount of land required to reach production targets.

4.3.3 Elephant conservation

We used an elephant migration index (EI) to evaluate the significance of the planning units (1 km^2) to elephant conservation by examining their landscape connectivity based on environmental suitability. Environmental suitability of elephants was calculated using a multi-scale species distribution model (SDM) that combines expert range maps and occurrence data. The model consists of two separate Isolation forest-based SDMs (Song & Estes, 2023) using polygon-based observations (expert range maps and census blocks) and point-based observations, which were then combined using Baye fusion. Data for the model was obtained from the African Elephant Database (African Elephant Specialist Group, 2023) for polygon-based maps and from Global Biodiversity Information Facility (GBIF.org, 2023) for presence-only occurrences. The model considers various environmental factors, such as vegetation index, groundwater, human impact (roads, settlements, agriculture), climate, topography, and landscape metrics. The map of environmental suitability and census blocks can be found in Figure D-3 in 4.5Appendix D.

We applied circuit theory via Circuitscape 5 (Hall et al., 2021) with pairwise mode to simulate the landscape connectivity. We utilized the environmental suitability obtained in the previous step as conductance in Circuitscape and census blocks of elephants as nodes. Specifically, 50% of the census blocks were randomly selected as nodes and the simulation was run 50 times to obtain spatial connectivity both within and outside of habitats. The average of all iterations' cumulated currents represents the landscape connectivity. The elephant migration index in a given planning unit i is then calculated by standardizing landscape connectivity (C_i):

$$EI_i = \max\left(0, min\left(1, \frac{C_i - C_{q1}}{C_{q1} - C_{q99}}\right)\right) \quad (4\text{-}5)$$

where C_{q1} is the 1% quantile and C_{q99} is the 99% quantile of C_i over the entire region. The usage of quantiles is to reduce the impacts of extreme values.

4.3.4 Biodiversity cost

To cover different aspects of evaluating biodiversity values and potential impacts (Crawford et al., 2021; Soto-Navarro et al., 2020), we included multiple biodiversity metrics calculated with different data resources and ecological criteria as indicators of biodiversity status at ecosystem and species levels (Table 4-1). At ecosystem level, we used CSIRO Biodiversity Habitat Index (BHI) as the evaluation of local habitat importance (Ferrier et al., 2004) and took GLOBIO Mean Species Abundance (MSA) and PREDICTS Biodiversity Intactness Index (BII) as measures of regional ecosystem intactness (Newbold et al., 2016; Schipper et al., 2020). BHI evaluates the impacts of habitat destruction, fragmentation, and degradation for a given spatial reporting unit (e.g. 1 km in this study) on retention of terrestrial biodiversity globally (Ferrier et al., 2004). It ranges from 0 to 1, with one being the most important for retention of habitat

supporting distinct assemblages of species (plants, vertebrates, and invertebrates). MSA is an indicator measuring local terrestrial biodiversity intactness, calculated by quantitative human pressure-impact relationships. Human pressures include land use, road disturbance, habitat fragmentation, hunting, nitrogen deposition, and climate change (Schipper et al., 2020). It ranges from 0 to 1, where 0 indicates all species are locally extinct, and 1 indicates the species assemblage is undisturbed. BII is also a measure of ecosystem intactness, describing the current average abundance of native species assemblage in a given geographical unit, relative to their reference populations with no human pressures (Newbold et al., 2016).

We calculated species richness and rarity-weighted richness as species-level biodiversity assessment metrics. Species richness is the number of species in a given geographical unit. Rarity-weighted species richness is the sum of the proportion of each species' range contained in a given geographical unit. Compared to species richness, it can effectively weight species richness by global range size to give species with smaller ranges greater weight. All metrics were compiled using the refined geographical ranges of terrestrial mammals (n = 354), birds (n = 1067), reptiles (n = 347), and amphibians (n = 178). We additionally weighted these metrics by IUCN Red List Category, global range size, and species indigenousness and then rescaled them to an index ranging between 0 and 1 (see details in 4.5Appendix D). We queried range maps from IUCN Red List for terrestrial mammals and amphibians (IUCN, 2017), from BirdLife International for birds (BirdLife International, 2022), and from GARD for reptiles (Roll et al., 2017; Roll & Meiri, 2022). To reduce the commission errors, we excluded areas with unsuitable habitat types and elevation values from the geographical range maps (Brooks et al., 2019; Jung et al., 2020) using information on habitat types and elevation limits of species provided by IUCN Red List (IUCN, 2017).

Following previous studies (Brooks et al., 2006; Soto-Navarro et al., 2020), we ensembled multiple biodiversity metrics into a proactive biodiversity index (BIp) as a composite biodiversity indicator. Areas with high BIp have high local biodiversity, intactness, and compositional similarity of habitat conditions across the broader area to support (or previously) a distinct assemblage of species, and thus need biodiversity protection. For each planning unit i, it is calculated as following (Soto-Navarro et al., 2020):

$$BIp_i = b_i \times c_i \quad (4\text{-}6)$$

$$b_i = S_i + E_i \quad (4\text{-}7)$$

where b_i indicates local biodiversity and intactness for a given planning unit i, c_i represents the compositional similarity of habitat condition in a given planning unit i to support (or previously) distinct natural assemblages of species. We used the arithmetic mean of species-level biodiversity (S_i) and ecosystem-level biodiversity intactness (E_i) to represent b_i, and directly used BHI to describe c_i. S_i is calculated as the geometric mean of overall species richness and rarity-weighted species richness averaging among four taxonomic groups (mammal, amphibian, bird, and reptile) in a given planning unit i. E_i is calculated as the arithmetic mean of MSA and BII for a given planning unit i. More details and maps of these variables can be found in section 4.5D.3 in 4.5Appendix D.

4.3.5 Carbon cost

We developed maps of above- and below-ground terrestrial carbon stocks (tonnes per hectare) in vegetation biomass and soil for Tanzania to estimate potential carbon costs from land conversion. To ensure the robustness of the carbon stock estimation, we combined multiple

publicly available datasets (FAO, 2018; Lin et al., 2022; Poggio et al., 2021; Santoro & Cartus, 2021; Veiga & Balzter, 2021), which were selected based on reference year, spatial resolution, and accuracy. We determined the above-ground vegetation biomass by a pixel-wise uncertainty-weighted average of the NCEO Africa Aboveground Biomass map (Veiga & Balzter, 2021) and ESA CCI biomass map (Santoro & Cartus, 2021). The below-ground biomass was estimated from above-ground biomass using land cover-specified root-shoot ratios. We used 0.531 as the root-shoot ratio for Miombo woodland savanna (Mokany et al., 2006; Ryan et al., 2011), 1.887 for tropical/subtropical grassland (including other sparse vegetation regions such as bareland and wetland), and 0.205/0.235 for tropical/subtropical forest with shoot biomass lower/higher than 125 Mg ha^{-1} (Mokany et al., 2006). The total vegetation biomass was then multiplied by 0.49 (Xu et al., 2021) to convert to carbon.

Similarly, we estimated soil carbon density in the top 1 m based on soil profile maps of GSOCmap (FAO, 2018) and SoilGrids (Poggio et al., 2021). More specifically, we averaged the soil organic carbon stocks in the top 30 cm by averaging the maps made by GSOCmap and SoilGrids. We then calculated the carbon storage from 1 m up to 30 cm with the soil organic carbon content and bulk density of layers 30-60 cm and 60 – 100 cm made by SoilGrids (Poggio et al., 2021). The total soil organic carbon stocks in the top 1 m are the sum of carbon stocks in the top 30 cm and 30 – 100 cm. The figures in Figure D-6 (4.5Appendix D) illustrate these inputs generated for carbon cost calculation.

Finally, we added 100% of the vegetation carbon storage and 25% of the soil organic carbon storage (Estes et al., 2016) and multiplied by actual conversion area within each planning unit as the combined total carbon loss due to land conversion.

4.3.6 Trade-off model structure

To target the areas for agricultural expansion in Tanzania for an optimal solution to maximize benefits towards agricultural production value and ecological maintenance, attainable yield (Ya), transport time (T), elephant migration index (EI), proactive biodiversity index (BIp), and carbon loss (Cl) for each planning unit were used as decision-making factors. We first express the different objectives as efficiencies, e.g. biodiversity loss per ton of crop yield, and normalized each factor to ranging from 0 to 1, with 1 being the highest value, and calculated the score of suitability (S) for each planning unit i as follows:

$$S_i = a \cdot Ya_i + b \cdot (1 - T_i) + c \cdot (1 - EI_i) + d \cdot (1 - BIp_i) + e \cdot (1 - Cl_i) \quad (4\text{-}8)$$

where weights a, b, c, d, and e represent the weights are placed on each decision-making factor, expressed as a decimal between 0 and 1, where values closer to 1 signifies stronger preferences, and the sum of all weights is equal to 1. Weights a and b together represent the preference of agricultural benefit. The suitability of a planning unit for agricultural expansion increases with a higher value of S_i. All planning units were then ranked by S_i and selected in descending order until the cumulative production reaches production target.

In this study, we evaluated two scenarios based on the maximum area in each planning unit to be converted to arable land: 60% and 80%. Under both scenarios, the planning units that surpass the maximum allowed area for current cropland will still retain their cropland status. In the 60% scenario, each planning unit will allocate a larger proportion for non-agricultural purposes, resulting in a more dispersed distribution of cropland, which would be advantageous for small-sized species. Conversely, in the 80% scenario, each planning unit will allocate a smaller proportion for non-agricultural purposes, resulting in a more compact distribution of

cropland, which would leave more undisturbed open areas to benefit large-sized migratory species, such as African elephants. We excluded areas of PAs as not allowed farming region for all solutions to protect species' habitats.

4.4 Results

4.4.1 Status and effectiveness of protected areas (PAs) in Tanzania

In Tanzania, about 40% of the terrestrial area is designed as protected areas (PAs) since the establishment of the first national park in 1951 (Riggio et al., 2019). This country has a high number of PAs, contributing to the impressive regional conservation trend. The areas within PAs have significantly higher values of biodiversity index, elephant migration index, and carbon density compared to areas outside of PAs (Figure 4-1 & Table 4-1). Thus, these protected areas effectively protect natural habitats, acting as a land-sparing practice. Despite the high number of PAs in Tanzania, a significant portion (> 8%) of the protected area in Tanzania has been converted to agriculture, accounting for 15% of the total agricultural area (Table 4-1). Additionally, in certain regions, such as around the Moyowosi Game Reserve in northwest Tanzania, the western boundary of Serengeti National Park, and southwestern boundary of Tarangire National Park in northern Tanzania (Figure 4-1), extensive anthropogenic land cover (both agriculture and settlement) has occupied most of the areas outside of PAs, leading to the loss of ecologically meaningful buffer zones.

Table 4-1. Agricultural and ecological statistics inside and outside of protected areas (PAs) in Tanzania. Plant area is the overall area in hectares. Other values, proactive biodiversity index (BIp), carbon density, and elephant migration index (EI), use the mean values.

Location	Plant area ($ha \times 10^6$)	Mean proactive biodiversity index (0-1)	Mean carbon density (t/ha)	Mean elephant migration index (0-1)
Outside of PAs	19.86	0.41±0.1	49.07±23.3	0.33±0.23
Inside of PAs	3.66	0.56±0.1	59.71±30.7	0.48±0.25

4.4.2 Productive benefits and ecological costs for agricultural development in Tanzania

The attainable production gain can be achieved through either intensification of agricultural practices in existing croplands or expansion into non-farm regions (section 4.3.2). The expansion of agriculture can involve expanding into previously undisturbed land or expanding existing farms. Our findings indicate that expansion into new regions will result in a higher production gain in the near future, with the southeastern coastal region being particularly conducive to growing additional crops to meet rising food demand (Figure 4-2a). However, it is important to note that expansion into new areas will result in considerably higher environmental costs, including increased conservation and carbon costs, as compared to intensification in existing croplands (Figure 4-2c&d). These trade-offs highlight the necessity for optimizing land for agricultural expansion. Our analysis reveals that the spatial distribution of impacts on biodiversity and carbon costs of land conversion is negatively associated with the extent of existing cropland (Figure 4-1 & Figure 4-2c-d). This is because areas with established croplands have already undergone human disturbance and have less risk of causing ecological crises. However, the distribution of the impact on landscape connectivity is not as closely related to the extent of current cropland (Figure 4-2e). This may indicate the lack of consideration given to

landscape connectivity in land use management in the past. Additionally, expanding agriculture into new areas will result in lower accessibility to trade markets due to the lack of infrastructure, such as roads (Figure 4-2f). As a result, alternative approaches to increasing crop yields, such as implementing advanced farming management practices (e.g., utilizing fertilizers) or expanding cultivated areas adjacent to existing croplands (e.g., increasing field size), may be more profitable.

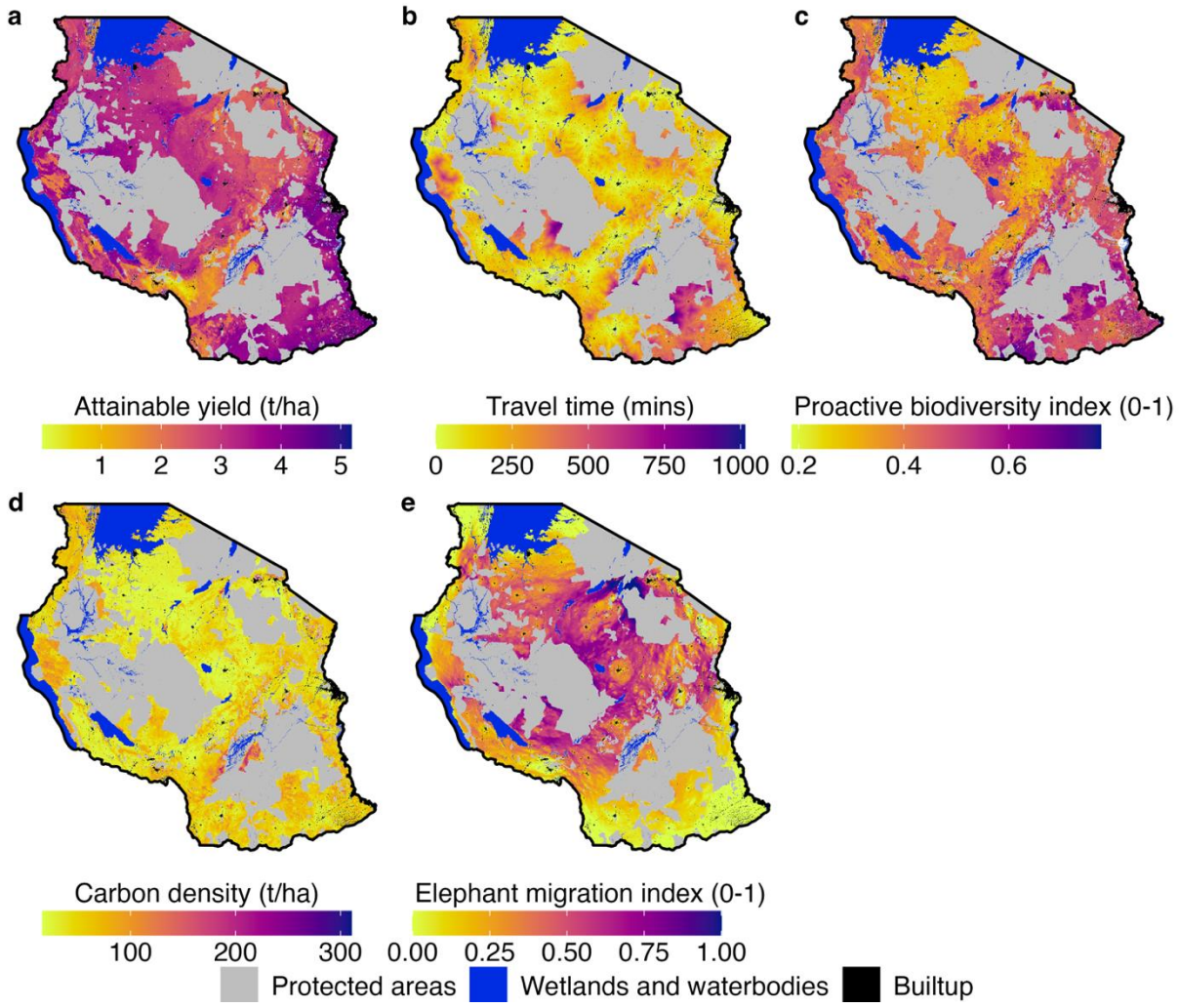

Figure 4-2. Decision-making factors of each planning unit in land allocation analysis: attainable yield (a), travel time (b), proactive biodiversity index (c), carbon density (d), and elephant migration index (e).

4.4.3 Contribution of landscape connectivity in land management

In this study, we assessed the effectiveness of different criteria for lowering ecological costs in new agricultural land allocation. We analyzed three solutions under the 60% scenario (section 4.3.6), based on various agricultural and ecological goals (Figure 4-3). When only considering agricultural benefits (both yield and travel time), farmland was allocated widely across the whole region in the land with highest production benefit, resulting in higher number of undisturbed units selected for agriculture expansion compared to other solutions (Figure 4-3a & Figure 4-2a-b). As a result, this solution led to more expansion into non-cultivated areas, such as the southeastern corner of Tanzania and northeastern coastal regions, causing high carbon and biodiversity loss (Figure 4-3). When biodiversity and carbon cost were added to the evaluation, more cropland expansion was predicted to happen in current farming areas (Figure 4-1 & Figure 4-3b), leading to lower carbon and biodiversity loss compared to the agriculture-only solution.

The result suggests that agricultural development outside of PAs in Tanzania may not lead to significant biodiversity loss (Figure 4-2a & Figure 4-3), likely due to the effective protection provided by long-standing PAs that have stabilized the landscape. However, with increasing human activities, protected areas can no longer ensure sufficient landscape connectivity for long-distance wildlife migration (e.g. by African elephants). Therefore, some areas outside of protected areas have high ecological costs for agricultural development if considering landscape connectivity. Including habitat connectivity in the ecological cost evaluation results in a more compacted and fragmented agricultural allocation with multiple centers (Figure 4-3c) to spare enough space for species migration. Such adjustment requires

necessary expansion into new areas, leading to inevitable higher carbon loss (Figure 4-3) compared to the solution without considering landscape connectivity.

In general, incorporating ecological costs into agricultural land allocation decreases the expansion to undisturbed areas and reduces the ecological loss, despite the different advantages of each solution (Figure 4-3).

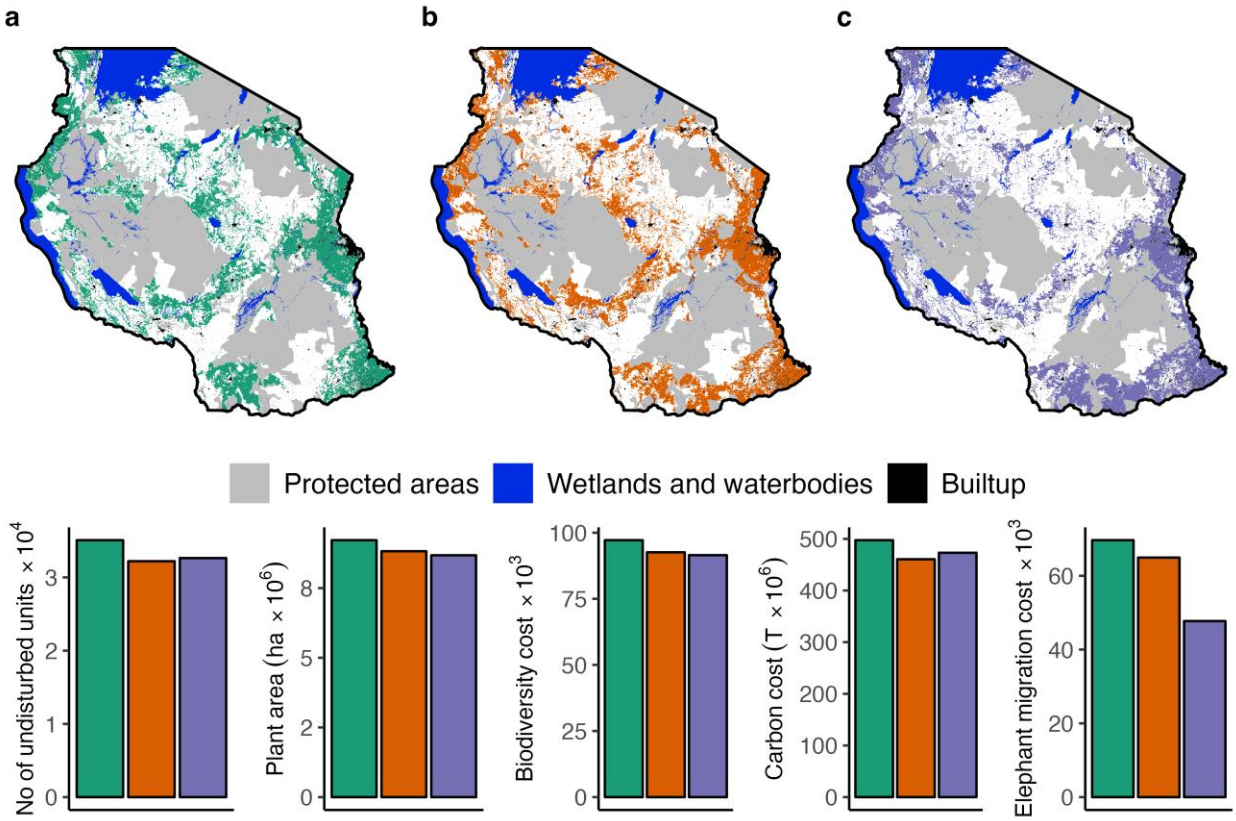

Figure 4-3. New selected agricultural areas and ecological costs under different solutions: (a) only considering agricultural benefits (attainable yield and transport time), (b) considering both agricultural benefits and ecological costs (biodiversity cost and carbon cost), and (c) considering agricultural benefits, ecological costs (biodiversity cost, carbon cost, and elephant migration cost.

4.4.4 Significance of agricultural intensification

The findings of our study reveal comparable spatial land allocation patterns for two scenarios (60% and 80%, section 4.3.6). Notably, the 60% scenario exhibited a 23% more selection of planning units than the 80% scenario, yet only yielded a marginal increase of 2.5% (~200,000 ha) in farmland area (see Table 4-2). These results suggest that the amount of farmland required to meet food demand is primarily determined by the land's productivity, whereas the distribution of areas and ecological costs could vary according to the underlying criteria.

Table 4-2. New allocated cropland and ecological costs under different scenarios.

Areas and costs	Scenario*			
	80%	60%	80% with half yield gain	60% with half yield gain
No of planning units to expand	168558	206646	254015	321758
Current cropland to expand (ha × 10^6)	6.15	5.30	8.91	7.75
New land to expand (ha × 10^6)	2.31	3.37	4.67	6.49
Biodiversity cost × 10^4	7.09	9.15	10.95	14.50
Carbon cost (T × 10^6)	428.40	473.04	724.48	816.47
Elephant migration cost × 10^3	34.29	47.73	61.68	93.59
Mean travel time (hours)	2.40	2.88	2.75	3.33

*60% and 80% indicate the allowed maximum percentage of each planning unit (100 ha) to expand for cultivation. Planning units currently with cultivated areas higher than #% would not be reduced.

In the main analysis, we assumed that the yield gap between current and projected attainable yield would be fully closed. Under this assumption, a considerable amount of land outside of protected areas (PAs) remained open, even when 60% of each planning unit was allowed for agriculture (Figure 4-4b). We further explored the model by considering scenarios where only half of the yield gap was closed. Our results revealed that compared to the case of complete yield gap closure, the required cropland area and the associated ecological loss nearly

doubled (see Table 4-2). Particularly, under the scenario that 60% of each planning unit can be cultivated, nearly all the remaining land outside of PAs was allocated for agricultural use, and as a result, the PAs became highly isolated from each other (Figure 4-4d). This isolation could lead to indirect negative impacts on the species residing in each PAs that lack connections with other groups.

Hence, implementing agricultural practices to close the yield gap is decisive in addressing the increasing demand for food while ensure conservation sustainability. Spatial planning can aid in optimizing land use and minimizing ecological loss, but only if the attainable yield is sufficiently high to warrant making decisions about which land to spare.

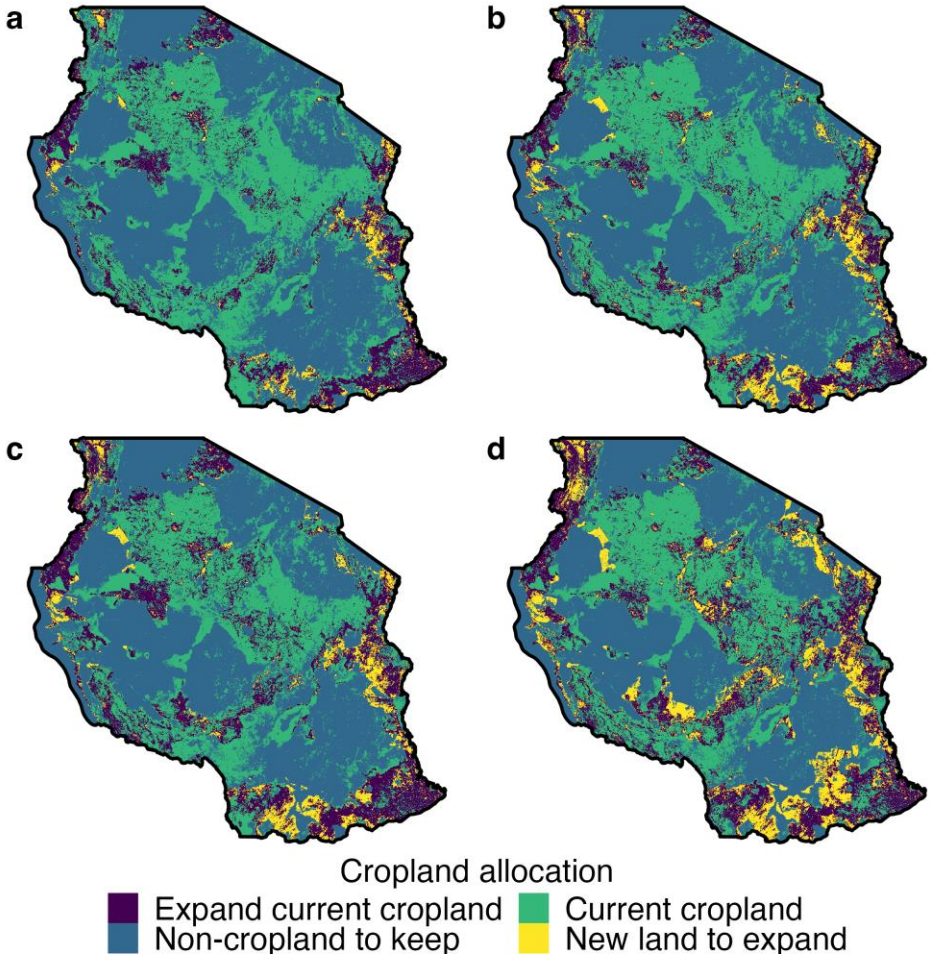

Figure 4-4. Comparison between current cropland and newly allocated cropland under 80% (a), 60% (b), 80% with half yield gap close (c), and 60% with half yield gap close (d) scenario. Current cropland includes all planning units with cultivated area. #% indicates the allowed maximum percentage of each planning unit (100 ha) to expand for cultivation. Planning units currently with cultivated areas higher than #% would not be reduced. Equal weights were set across all objectives.

4.4.5 Harmonize different ecological criteria

In order to achieve a balanced benefit among the three ecological objectives of biodiversity, carbon, and elephant conservation, we investigated the cost translation relationships between these factors by assigning different weights to them during decision-making (Eq. (4-8 &

section 4.3.6). Under the 60% scenario, we applied a fixed weight (25%) to agricultural benefit and distributed it equally attainable yield and transport time. We then adjusted the weights for biodiversity loss, carbon loss, and elephant migration loss between 0 and 75%. To allow for comparison between the values, we normalized the values of the three ecological costs between their minimum and maximum values. The standard deviation of the normalized cost values reflects the balance between the ecological objectives for each solution. As shown in Figure 4-5a, optimal solutions that maximize one objective inevitably led to high losses in the other objectives. Interestingly, increasing the cost of one objective exponentially reduces the loss of the other objectives. In other words, sacrificing a little of one objective can significantly benefit the other objectives when compared to maximizing this objective.

Our analysis shows that a weight of 17% for biodiversity, 28% for carbon, and 30% for elephant migration in land allocation decision-making achieves an equally balanced solution for all three ecological objectives (Figure 4-5a). The resulting land for agricultural expansion is depicted in Figure 4-5b, with a selection of 206754 planning units and an agricultural expansion of 8.7×10^6 ha. The ecological costs are 92.3×10^3 biodiversity loss, 475.4×10^6 tonnes carbon loss, and 45.0×10^3 elephant migration loss.

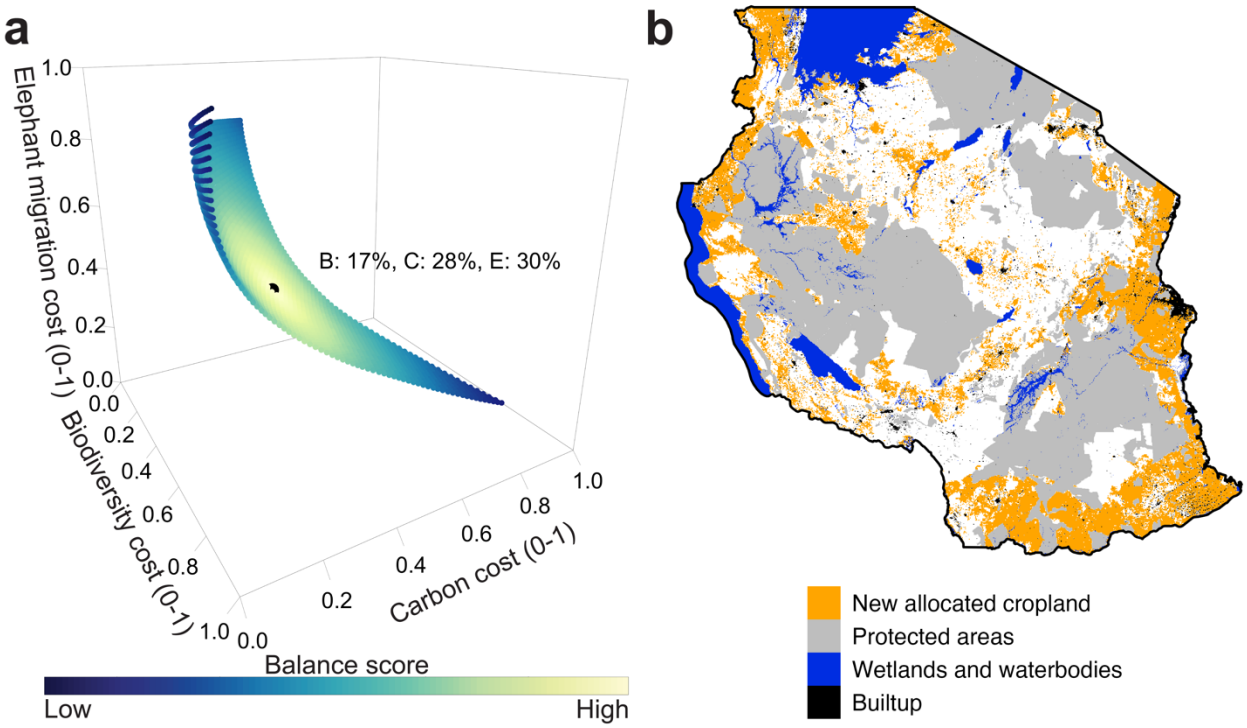

Figure 4-5. Tradeoff surface between biodiversity cost, carbon cost and elephant migration cost using different weights in the decision-making (a), and agricultural land expansion with the highest balance between ecological costs (b). The tradeoff surface was generated using a fixed weight (25%) for the agricultural benefit, which was equally split between attainable yield and transport time. The weight for each analyzed ecological factor ranges from 0 to 75%. The values of three ecological costs were normalized between their minimum and maximum values. The balance score is the standard deviation of three cost values.

4.5 Discussion

Integrating various objectives among stakeholders is a crucial step towards achieving evidence-based and sustainable design of agricultural systems, which is a major challenge of our time (Cassman & Grassini, 2020; Kremen, 2015; Kremen & Merenlender, 2018). Having high enough land productivity to meet food security demands is the prerequisite for achieving cost-efficient land allocation (Bonilla-Cedrez et al., 2021; Jayne et al., 2019; van Ittersum et al., 2016). Sub-Saharan African countries, including Tanzania, are known for their smallholder

farming systems, which exhibit persistently low productivity levels, with an average remaining yield gap of approximately 80% (van Ittersum et al., 2016). Recent increases in production have been primarily due to area expansion rather than yield gains (URT, 2021), resulting in a range of ecological impacts. As demonstrated in this study (Figure 4-4), boosting yields on existing cropland has the potential, and arguably is the most fundamental strategy (see section 4.4.4), to increase national production to meet food security needs while mitigating ecological costs (Ewers et al., 2009). However, reducing yield gaps in Tanzania is challenging, as the profitability and availability of agricultural practices, such as fertilizer usage, varies greatly across different locations (Palmas & Chamberlin, 2020; Senthilkumar et al., 2020). In our analysis, we assumed that yield gaps would be fully closed throughout the study area. However, to ensure a more robust analysis, attainable yields should be calibrated based on the spatially explicit profitability of crop management practices. Additionally, it's important to consider the potential negative environmental impacts of agricultural intensification, such as soil degradation, groundwater depletion, and pollution of air and water, which can directly and indirectly affect biodiversity. To accurately reconcile the goals of agriculture and ecosystem conservation, it is essential to quantitatively estimate these outcomes and incorporate it in the analysis.

As a country recognized for its rich biodiversity, Tanzania has established a robust network of protected areas (PAs), which provided efficient and effective protection for its ecosystems (Table 4-1 & section 4.4.1). While this land-sparing management strategy initially provided sufficient protection for Tanzania's biodiversity and carbon storage, the rapidly increasing agricultural land area has significantly impacted habitat connectivity, particularly for large migratory species like African savanna elephants (Riggio & Caro, 2017). The isolation of PAs has led to a reduction in effective size, hindered gene flow between populations, and

resulted in increased human-wildlife conflicts (Lohay et al., 2020). Urgent action is then required to mitigate this situation through appropriate management practices that maintain the necessary interconnection between species groups in different habitat fragments. In this study, we integrated the landscape connectivity of African savanna elephants into our workflow to optimize agricultural land allocation. Our results demonstrate that this approach can effectively mitigate the negative impacts of agricultural development on habitat connectivity and provide spatially explicit recommendations for land use planning (see section 4.4.3 & 4.4.5).

Our analysis in this study represents a primary attempt to take landscape connectivity into the analysis framework of reconciling agriculture and ecosystems. We used a widely used approach, circuit theory (Hall et al., 2021), to model and evaluate the spatial importance for maintaining landscape connectivity. However, current methods of evaluating connectivity could be improved in several ways. Quantifying the importance of habitat connectivity robustly is crucial for designing effective conservation strategies. For instance, the loss of corridors can potentially reduce local passage habitats, disrupt the reciprocal dependability between resident and passing animals, and even affect population dynamics in far-off habitats through the butterfly effect (Donald & Evans, 2006). Conversely, the restoration of wildlife corridors can positively impact remote habitats (Perino et al., 2019). To accurately evaluate the importance of land for supporting habitat connectivity, it is necessary to obtain ground-based observations and develop both ecological mechanisms and computational models (Allen & Singh, 2016; Bolger et al., 2008; Doherty & Driscoll, 2018).

In landscape ecology analysis, scaling is an important consideration (Wu, 2013). Finer spatial resolutions can provide better insights into local impacts of human activities and habitat features that can change over short distances (Mokany et al., 2020). However, these finer

resolutions may also introduce more errors and uncertainties. In this study, we used planning units of 1 km^2 (100 ha) for land use management planning and ecological cost evaluation. We assumed that each planning unit can be partially cultivated, and used a 4.77 m resolution land cover map (Song et al., 2023) to obtain current farmland area and calculate areas for expansion. This approach can improve the precision of crop production and carbon loss estimation, while also evaluating conservation effectiveness within a necessary background window.

This study demonstrates that it is possible to achieve ideal solutions that maximize various objectives, albeit at the cost of sacrificing other objectives. It provides a baseline workflow for land use planning that can satisfy the demands of various stakeholders. The approach can identify areas suitable for agricultural expansion and redistribute current croplands to prevent the exacerbation of existing impacts. Furthermore, our results suggest that reducing the level of consideration given to a specific ecological service that stakeholders are willing to accept could significantly benefit other stakeholders' demands. In essence, a solution that reconciles different demands could be more beneficial when all stakeholders' objectives are considered together and with careful management. To ensure the legitimacy and usefulness of this type of analysis, several critical components must be considered. Among these, the quality and representation of data are indispensable (L. Estes et al., 2018; Estes et al., 2016). Given the sensitivity of the analysis to errors in the input data used for calculations, continued efforts to improve data quality are necessary for long-term use of this method. Additionally, incorporating uncertainties within the analysis workflow can improve the robustness of the results.

Acknowledgments

This work was supported by the Future Investigators in NASA Earth and Space Science and Technology (FINESST) program (award number: 80NSSC20K1640).

References

African Elephant Specialist Group. (2023). *African Elephant Database*. IUCN SSC. https://africanelephantdatabase.org/

Allen, A. M., & Singh, N. J. (2016). Linking movement ecology with wildlife management and conservation. *Frontiers in Ecology and Evolution*, *3*, 155.

BirdLife International. (2022). *BirdLife International and handbook of the birds of the world (2022) Bird species distribution maps of the world* [Data set]. http://datazone.birdlife.org/species/requestdis

Bolger, D. T., Newmark, W. D., Morrison, T. A., & Doak, D. F. (2008). The need for integrative approaches to understand and conserve migratory ungulates. *Ecology Letters*, *11*(1), 63–77.

Bonilla-Cedrez, C., Chamberlin, J., & Hijmans, R. J. (2021). Fertilizer and grain prices constrain food production in sub-Saharan Africa. *Nature Food*, *2*(10), 766–772.

Brancalion, P. H. S., Niamir, A., Broadbent, E., Crouzeilles, R., Barros, F. S. M., Almeyda Zambrano, A. M., Baccini, A., Aronson, J., Goetz, S., Reid, J. L., Strassburg, B. B. N., Wilson, S., & Chazdon, R. L. (2019). Global restoration opportunities in tropical rainforest landscapes. *Science Advances*, *5*(7), eaav3223. https://doi.org/10.1126/sciadv.aav3223

Brooks, T. M., Mittermeier, R. A., Da Fonseca, G. A., Gerlach, J., Hoffmann, M., Lamoreux, J. F., Mittermeier, C. G., Pilgrim, J. D., & Rodrigues, A. S. (2006). Global biodiversity conservation priorities. *Science*, *313*(5783), 58–61.

Brooks, T. M., Pimm, S. L., Akçakaya, H. R., Buchanan, G. M., Butchart, S. H. M., Foden, W., Hilton-Taylor, C., Hoffmann, M., Jenkins, C. N., Joppa, L., Li, B. V., Menon, V., Ocampo-Peñuela, N., & Rondinini, C. (2019). Measuring Terrestrial Area of Habitat (AOH) and Its Utility for the IUCN Red List. *Trends in Ecology & Evolution*, *34*(11), 977–986. https://doi.org/10.1016/j.tree.2019.06.009

Cardillo, M., Mace, G. M., Jones, K. E., Bielby, J., Bininda-Emonds, O. R. P., Sechrest, W., Orme, C. D. L., & Purvis, A. (2005). Multiple Causes of High Extinction Risk in Large Mammal Species. *Science*, *309*(5738), 1239–1241. https://doi.org/10.1126/science.1116030

Cassman, K. G., & Grassini, P. (2020). A global perspective on sustainable intensification research. *Nature Sustainability*, *3*(4), 262–268.

Clauzel, C., & Godet, C. (2020). Combining spatial modeling tools and biological data for improved multispecies assessment in restoration areas. *Biological Conservation*, *250*, 108713.

Crawford, C. L., Estes, L. D., Searchinger, T. D., & Wilcove, D. S. (2021). Consequences of underexplored variation in biodiversity indices used for land-use prioritization. *Ecological Applications*, *31*(7). https://doi.org/10.1002/eap.2396

Dickman, A. J. (2010). Complexities of conflict: The importance of considering social factors for effectively resolving human-wildlife conflict: Social factors affecting human-wildlife conflict resolution. *Animal Conservation*, *13*(5), 458–466. https://doi.org/10.1111/j.1469-1795.2010.00368.x

Doggart, N., Morgan-Brown, T., Lyimo, E., Mbilinyi, B., Meshack, C. K., Sallu, S. M., & Spracklen, D. V. (2020). Agriculture is the main driver of deforestation in Tanzania. *Environmental Research Letters*, *15*(3), 034028. https://doi.org/10.1088/1748-9326/ab6b35

Doherty, T. S., & Driscoll, D. A. (2018). Coupling movement and landscape ecology for animal conservation in production landscapes. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1870), 20172272. https://doi.org/10.1098/rspb.2017.2272

Donald, P. F., & Evans, A. D. (2006). Habitat connectivity and matrix restoration: The wider implications of agri-environment schemes: Habitat connectivity and matrix restoration. *Journal of Applied Ecology*, *43*(2), 209–218. https://doi.org/10.1111/j.1365-2664.2006.01146.x

Dudley, N., & Alexander, S. (2017). Agriculture and biodiversity: A review. *Biodiversity*, *18*(2–3), 45–49. https://doi.org/10.1080/14888386.2017.1351892

Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A. Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., … E. Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Elith, J., & Leathwick, J. R. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Estes, L., Chen, P., Debats, S., Evans, T., Ferreira, S., Kuemmerle, T., Ragazzo, G., Sheffield, J., Wolf, A., Wood, E., & Caylor, K. (2018). A large-area, spatially continuous assessment of land cover map error and its impact on downstream analyses. *Global Change Biology*, *24*(1), 322–337. https://doi.org/10.1111/gcb.13904

Estes, L. D., Searchinger, T., Spiegel, M., Tian, D., Sichinga, S., Mwale, M., Kehoe, L., Kuemmerle, T., Berven, A., Chaney, N., Sheffield, J., Wood, E. F., & Caylor, K. K. (2016). Reconciling agriculture, carbon and biodiversity in a savannah transformation frontier. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1703), 20150316. https://doi.org/10.1098/rstb.2015.0316

Ewers, R. M., Scharlemann, J. P. W., Balmford, A., & Green, R. E. (2009). Do increases in agricultural yield spare land for nature? *Global Change Biology*, *15*(7), 1716–1726. https://doi.org/10.1111/j.1365-2486.2009.01849.x

FAO. (2018). *Global Soil Organic Carbon Map (GSOCmap) Technical Report*. FAO.

Fastré, C., Possingham, H. P., Strubbe, D., & Matthysen, E. (2020). Identifying trade-offs between biodiversity conservation and ecosystem services delivery for land-use decisions. *Scientific Reports*, *10*(1), 7971. https://doi.org/10.1038/s41598-020-64668-z

Ferrier, S., Powell, G. V. N., Richardson, K. S., Manion, G., Overton, J. M., Allnutt, T. F., Cameron, S. E., Mantle, K., Burgess, N. D., Faith, D. P., Lamoreux, J. F., Kier, G., Hijmans, R. J., Funk, V. A., Cassis, G. A., Fisher, B. L., Flemons, P., Lees, D., Lovett, J. C., & Van Rompaey, R. S. A. R. (2004). Mapping More of Terrestrial Biodiversity for Global Conservation Assessment. *BioScience*, *54*(12), 1101. https://doi.org/10.1641/0006-3568(2004)054[1101:MMOTBF]2.0.CO;2

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302–4315. https://doi.org/10.1002/joc.5086

Fischer, G., Nachtergaele, F. O., van Velthuizen, H., Chiozza, F., Francheschini, G., Henry, M., Muchoney, D., & Tramberend, S. (2021). *Global agro-ecological zones (gaez v4)-model documentation*.

GBIF.org. (2023). *GBIF Occurrence Download*. https://doi.org/10.15468/dl.scpfwj

Grogan, D., Frolking, S., Wisser, D., Prusevich, A., & Glidden, S. (2022). Global gridded crop harvested area, production, yield, and monthly physical area data circa 2015. *Scientific Data*, *9*(1), 15. https://doi.org/10.1038/s41597-021-01115-2

Haddad, N. M., Brudvig, L. A., Clobert, J., Davies, K. F., Gonzalez, A., Holt, R. D., Lovejoy, T. E., Sexton, J. O., Austin, M. P., Collins, C. D., Cook, W. M., Damschen, E. I., Ewers, R. M., Foster, B. L., Jenkins, C. N., King, A. J., Laurance, W. F., Levey, D. J., Margules, C. R., … Townshend, J. R. (2015). Habitat fragmentation and its lasting impact on Earth's ecosystems. *Science Advances*, *1*(2), e1500052. https://doi.org/10.1126/sciadv.1500052

Hall, K. R., Anantharaman, R., Landau, V. A., Clark, M., Dickson, B. G., Jones, A., Platt, J., Edelman, A., & Shah, V. B. (2021). Circuitscape in Julia: Empowering Dynamic Approaches to Connectivity Assessment. *Land*, *10*(3), 301. https://doi.org/10.3390/land10030301

Hoare, R. (2015). Lessons From 20 Years of Human–Elephant Conflict Mitigation in Africa. *Human Dimensions of Wildlife*, *20*(4), 289–295. https://doi.org/10.1080/10871209.2015.1005855

Hofman, M. P. G., Hayward, M. W., Kelly, M. J., & Balkenhol, N. (2018). Enhancing conservation network design with graph-theory and a measure of protected area effectiveness: Refining wildlife corridors in Belize, Central America. *Landscape and Urban Planning*, *178*, 51–59. https://doi.org/10.1016/j.landurbplan.2018.05.013

IUCN. (2022). *The IUCN red list of threatened species. Version 2022-2*.

Jayne, T. S., Snapp, S., Place, F., & Sitko, N. (2019). Sustainable agricultural intensification in an era of rural transformation in Africa. *Global Food Security*, *20*, 105–113.

Jung, M., Dahal, P. R., Butchart, S. H. M., Donald, P. F., De Lamo, X., Lesiv, M., Kapos, V., Rondinini, C., & Visconti, P. (2020). A global map of terrestrial habitat types. *Scientific Data*, *7*(1), 256. https://doi.org/10.1038/s41597-020-00599-8

Kremen, C. (2015). Reframing the land-sparing/land-sharing debate for biodiversity conservation. *Annals of the New York Academy of Sciences*, *1355*(1), 52–76.

Kremen, C., & Merenlender, A. M. (2018). Landscapes that work for biodiversity and people. *Science*, *362*(6412), eaau6020.

Lin, Z., Dai, Y., Mishra, U., Wang, G., Shangguan, W., Zhang, W., & Qin, Z. (2022). On the magnitude and uncertainties of global and regional soil organic carbon: A comparative analysis using multiple estimates. *Earth System Science Data Discussions*, 1–24.

Lobell, D. B., Thau, D., Seifert, C., Engle, E., & Little, B. (2015). A scalable satellite-based crop yield mapper. *Remote Sensing of Environment*, *164*, 324–333. https://doi.org/10.1016/j.rse.2015.04.021

Lohay, G. G., Weathers, T. C., Estes, A. B., McGrath, B. C., & Cavener, D. R. (2020). Genetic connectivity and population structure of African savanna elephants (*Loxodonta*

africana) in Tanzania. *Ecology and Evolution*, *10*(20), 11069–11089. https://doi.org/10.1002/ece3.6728

Luhunga, P. M., Kijazi, A. L., Chang'a, L., Kondowe, A., Ng'ongolo, H., & Mtongori, H. (2018). Climate Change Projections for Tanzania Based on High-Resolution Regional Climate Models From the Coordinated Regional Climate Downscaling Experiment (CORDEX)-Africa. *Frontiers in Environmental Science*, *6*, 122. https://doi.org/10.3389/fenvs.2018.00122

Mokany, K., Ferrier, S., Harwood, T. D., Ware, C., Di Marco, M., Grantham, H. S., Venter, O., Hoskins, A. J., & Watson, J. E. M. (2020). Reconciling global priorities for conserving biodiversity habitat. *Proceedings of the National Academy of Sciences*, *117*(18), 9906–9911. https://doi.org/10.1073/pnas.1918373117

Mokany, K., Raison, R. J., & Prokushkin, A. S. (2006). Critical analysis of root: Shoot ratios in terrestrial biomes: ROOT : SHOOT RATIOS IN TERRESTRIAL BIOMES. *Global Change Biology*, *12*(1), 84–96. https://doi.org/10.1111/j.1365-2486.2005.001043.x

Morin, E., Herrault, P.-A., Guinard, Y., Grandjean, F., & Bech, N. (2022). The promising combination of a remote sensing approach and landscape connectivity modelling at a fine scale in urban planning. *Ecological Indicators*, *139*, 108930. https://doi.org/10.1016/j.ecolind.2022.108930

Mueller, N. D., Gerber, J. S., Johnston, M., Ray, D. K., Ramankutty, N., & Foley, J. A. (2012). Closing yield gaps through nutrient and water management. *Nature*, *490*(7419), 254–257. https://doi.org/10.1038/nature11420

Myneni, R., Knyazikhin, Y., & Park, T. (2015). *MOD15A2H MODIS/Terra leaf area Index/FPAR 8-Day L4 global 500m SIN grid V006*. NASA EOSDIS Land Processes DAAC. https://lpdaac.usgs.gov/products/mod15a2hv006/

Newbold, T., Hudson, L. N., Arnell, A. P., Contu, S., De Palma, A., Ferrier, S., Hill, S. L. L., Hoskins, A. J., Lysenko, I., Phillips, H. R. P., Burton, V. J., Chng, C. W. T., Emerson, S., Gao, D., Pask-Hale, G., Hutton, J., Jung, M., Sanchez-Ortiz, K., Simmons, B. I., … Purvis, A. (2016). Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment. *Science*, *353*(6296), 288–291. https://doi.org/10.1126/science.aaf2201

Newbold, T., Hudson, L. N., Hill, S. L. L., Contu, S., Lysenko, I., Senior, R. A., Börger, L., Bennett, D. J., Choimes, A., Collen, B., Day, J., De Palma, A., Díaz, S., Echeverria-Londoño, S., Edgar, M. J., Feldman, A., Garon, M., Harrison, M. L. K., Alhusseini, T., … Purvis, A. (2015). Global effects of land use on local terrestrial biodiversity. *Nature*, *520*(7545), 45–50. https://doi.org/10.1038/nature14324

Palmas, S., & Chamberlin, J. (2020). Fertilizer profitability for smallholder maize farmers in Tanzania: A spatially-explicit ex ante analysis. *PLOS ONE*, *15*(9), e0239149. https://doi.org/10.1371/journal.pone.0239149

Perino, A., Pereira, H. M., Navarro, L. M., Fernández, N., Bullock, J. M., Ceauşu, S., Cortés-Avizanda, A., van Klink, R., Kuemmerle, T., Lomba, A., & others. (2019). Rewilding complex ecosystems. *Science*, *364*(6438), eaav5570.

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *SOIL*, *7*(1), 217–240. https://doi.org/10.5194/soil-7-217-2021

Préau, C., Tournebize, J., Lenormand, M., Alleaume, S., Boussada, V. G., & Luque, S. (2022). Habitat connectivity in agricultural landscapes improving multi-functionality of constructed wetlands as nature-based solutions. *Ecological Engineering*, *182*, 106725. https://doi.org/10.1016/j.ecoleng.2022.106725

Rands, M. R. W., Adams, W. M., Bennun, L., Butchart, S. H. M., Clements, A., Coomes, D., Entwistle, A., Hodge, I., Kapos, V., Scharlemann, J. P. W., Sutherland, W. J., & Vira, B. (2010). Biodiversity Conservation: Challenges Beyond 2010. *Science*, *329*(5997), 1298–1303. https://doi.org/10.1126/science.1189138

Riggio, J., & Caro, T. (2017). Structural connectivity at a national scale: Wildlife corridors in Tanzania. *PLOS ONE*, *12*(11), e0187407. https://doi.org/10.1371/journal.pone.0187407

Riggio, J., Jacobson, A. P., Hijmans, R. J., & Caro, T. (2019). How effective are the protected areas of East Africa? *Global Ecology and Conservation*, *17*, e00573. https://doi.org/10.1016/j.gecco.2019.e00573

Ripple, W. J., Newsome, T. M., Wolf, C., Dirzo, R., Everatt, K. T., Galetti, M., Hayward, M. W., Kerley, G. I. H., Levi, T., Lindsey, P. A., Macdonald, D. W., Malhi, Y., Painter, L. E., Sandom, C. J., Terborgh, J., & Van Valkenburgh, B. (2015). Collapse of the world's largest herbivores. *Science Advances*, *1*(4), e1400103. https://doi.org/10.1126/sciadv.1400103

Roll, U., Feldman, A., Novosolov, M., Allison, A., Bauer, A. M., Bernard, R., Böhm, M., Castro-Herrera, F., Chirio, L., Collen, B., Colli, G. R., Dabool, L., Das, I., Doan, T. M., Grismer, L. L., Hoogmoed, M., Itescu, Y., Kraus, F., LeBreton, M., … Meiri, S. (2017). The global distribution of tetrapods reveals a need for targeted reptile conservation. *Nature Ecology & Evolution*, *1*(11), 1677–1682. https://doi.org/10.1038/s41559-017-0332-2

Roll, U., & Meiri, S. (2022). *GARD 1.7—Updated global distributions for all terrestrial reptiles* [Data set]. Dryad. https://doi.org/10.5061/dryad.9cnp5hqmb

Rowhani, P., Lobell, D. B., Linderman, M., & Ramankutty, N. (2011). Climate variability and crop production in Tanzania. *Agricultural and Forest Meteorology*, *151*(4), 449–460. https://doi.org/10.1016/j.agrformet.2010.12.002

Running, S., Mu, Q., & Zhao, M. (2015). *MOD17A2H MODIS/terra gross primary productivity 8-day L4 global 500m SIN grid V006*. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MOD17A2H.006

Running, S., Mu, Q., & Zhao, M. (2021). *MOD16A2 MODIS/Terra Net Evapotranspiration 8-Day L4 Global 500m SIN Grid v006*. NASA EOSDIS Land Processes DAAC. https://lpdaac.usgs.gov/products/mod16a2v006/

Ryan, C. M., Williams, M., & Grace, J. (2011). Above- and Belowground Carbon Stocks in a Miombo Woodland Landscape of Mozambique: Carbon Stocks in a Miombo Woodland. *Biotropica*, *43*(4), 423–432. https://doi.org/10.1111/j.1744-7429.2010.00713.x

Santoro, M., & Cartus, O. (2021). ESA biomass climate change initiative (Biomass_cci): Global datasets of forest above-ground biomass for the years 2010, 2017 and 2018, v2. *Cent. Environ. Data Anal*.

Schipper, A. M., Hilbers, J. P., Meijer, J. R., Antão, L. H., Benítez-López, A., Jonge, M. M. J., Leemans, L. H., Scheper, E., Alkemade, R., Doelman, J. C., Mylius, S., Stehfest, E., Vuuren, D. P., Zeist, W., & Huijbregts, M. A. J. (2020). Projecting terrestrial biodiversity

intactness with GLOBIO 4. *Global Change Biology*, *26*(2), 760–771. https://doi.org/10.1111/gcb.14848

Senthilkumar, K., Rodenburg, J., Dieng, I., Vandamme, E., Sillo, F. S., Johnson, J., Rajaona, A., Ramarolahy, J. A., Gasore, R., Abera, B. B., Kajiru, G. J., Mghase, J., Lamo, J., Rabeson, R., & Saito, K. (2020). Quantifying rice yield gaps and their causes in Eastern and Southern Africa. *Journal of Agronomy and Crop Science*, *206*(4), 478–490. https://doi.org/10.1111/jac.12417

Shaffer, L. J., Khadka, K. K., Van Den Hoek, J., & Naithani, K. J. (2019). Human-Elephant Conflict: A Review of Current Management Strategies and Future Directions. *Frontiers in Ecology and Evolution*, *6*, 235. https://doi.org/10.3389/fevo.2018.00235

Song, L., Estes, A. B., & Estes, L. D. (2023). A super-ensemble approach to map land cover types with high resolution over data-sparse African savanna landscapes. *International Journal of Applied Earth Observation and Geoinformation*, *116*, 103152. https://doi.org/10.1016/j.jag.2022.103152

Song, L., & Estes, L. (2023). ITSDM: Isolation forest-based presence-only species distribution modelling and explanation in R. *Methods in Ecology and Evolution*, 2041-210X.14067. https://doi.org/10.1111/2041-210X.14067

Soto-Navarro, C., Ravilious, C., Arnell, A., de Lamo, X., Harfoot, M., Hill, S. L. L., Wearn, O. R., Santoro, M., Bouvet, A., Mermoz, S., Le Toan, T., Xia, J., Liu, S., Yuan, W., Spawn, S. A., Gibbs, H. K., Ferrier, S., Harwood, T., Alkemade, R., … Kapos, V. (2020). Mapping co-benefits for carbon storage and biodiversity to inform conservation policy and action. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1794), 20190128. https://doi.org/10.1098/rstb.2019.0128

Strassburg, B. B. N., Beyer, H. L., Crouzeilles, R., Iribarrem, A., Barros, F., de Siqueira, M. F., Sánchez-Tapia, A., Balmford, A., Sansevero, J. B. B., Brancalion, P. H. S., Broadbent, E. N., Chazdon, R. L., Filho, A. O., Gardner, T. A., Gordon, A., Latawiec, A., Loyola, R., Metzger, J. P., Mills, M., … Uriarte, M. (2018). Strategic approaches to restoring ecosystems can triple conservation gains and halve costs. *Nature Ecology & Evolution*, *3*(1), 62–70. https://doi.org/10.1038/s41559-018-0743-8

Tanzania National Bureau of Statistics. (2021). *Tanzania in Figures 2021*. Tanzania National Bureau of Statistics. https://www.nbs.go.tz/index.php/en/tanzania-in-figures/784-tanzania-in-figures-2021

Tarabon, S., Bergès, L., Dutoit, T., & Isselin-Nondedeu, F. (2019). Environmental impact assessment of development projects improved by merging species distribution and habitat connectivity modelling. *Journal of Environmental Management*, *241*, 439–449. https://doi.org/10.1016/j.jenvman.2019.02.031

Tilman, D., Clark, M., Williams, D. R., Kimmel, K., Polasky, S., & Packer, C. (2017). Future threats to biodiversity and pathways to their prevention. *Nature*, *546*(7656), 73–81. https://doi.org/10.1038/nature22900

Travers, H., Selinske, M., Nuno, A., Serban, A., Mancini, F., Barychka, T., Bush, E., Rasolofoson, R. A., Watson, J. E. M., & Milner-Gulland, E. J. (2019). A manifesto for predictive conservation. *Biological Conservation*, *237*, 12–18. https://doi.org/10.1016/j.biocon.2019.05.059

URT. (2021). *National sample census of agriculture 2019/20*. National report (Issue August).

van Ittersum, M. K., van Bussel, L. G. J., Wolf, J., Grassini, P., van Wart, J., Guilpart, N., Claessens, L., de Groot, H., Wiebe, K., Mason-D'Croz, D., Yang, H., Boogaard, H., van Oort, P. A. J., van Loon, M. P., Saito, K., Adimo, O., Adjei-Nsiah, S., Agali, A., Bala, A., … Cassman, K. G. (2016). Can sub-Saharan Africa feed itself? *Proceedings of the National Academy of Sciences*, *113*(52), 14964–14969. https://doi.org/10.1073/pnas.1610359113

Veiga, P. R., & Balzter, H. (2021). *Africa Aboveground Biomass map for 2017*. https://doi.org/10.25392/leicester.data.15060270.v1

Weiss, D. J., Nelson, A., Gibson, H. S., Temperley, W., Peedell, S., Lieber, A., Hancher, M., Poyart, E., Belchior, S., Fullman, N., Mappin, B., Dalrymple, U., Rozier, J., Lucas, T. C. D., Howes, R. E., Tusting, L. S., Kang, S. Y., Cameron, E., Bisanzio, D., … Gething, P. W. (2018). A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature*, *553*(7688), 333–336. https://doi.org/10.1038/nature25181

Williams, B. A., Grantham, H. S., Watson, J. E. M., Alvarez, S. J., Simmonds, J. S., Rogéliz, C. A., Da Silva, M., Forero-Medina, G., Etter, A., Nogales, J., Walschburger, T., Hyman, G., & Beyer, H. L. (2020). Minimising the loss of biodiversity and ecosystem services in an intact landscape under risk of rapid agricultural development. *Environmental Research Letters*, *15*(1), 014001. https://doi.org/10.1088/1748-9326/ab5ff7

Williams, D. R., Clark, M., Buchanan, G. M., Ficetola, G. F., Rondinini, C., & Tilman, D. (2020). Proactive conservation to prevent habitat losses to agricultural expansion. *Nature Sustainability*, *4*(4), 314–322. https://doi.org/10.1038/s41893-020-00656-5

Winkler, K., Fuchs, R., Rounsevell, M., & Herold, M. (2021). Global land use changes are four times greater than previously estimated. *Nature Communications*, *12*(1), 2501. https://doi.org/10.1038/s41467-021-22702-2

Wu, J. (2013). Key concepts and research topics in landscape ecology revisited: 30 years after the Allerton Park workshop. *Landscape Ecology*, *28*(1), 1–11. https://doi.org/10.1007/s10980-012-9836-y

Xu, L., Saatchi, S. S., Yang, Y., Yu, Y., Pongratz, J., Bloom, A. A., Bowman, K., Worden, J., Liu, J., Yin, Y., Domke, G., McRoberts, R. E., Woodall, C., Nabuurs, G.-J., de-Miguel, S., Keller, M., Harris, N., Maxwell, S., & Schimel, D. (2021). Changes in global terrestrial live biomass over the 21st century. *Science Advances*, *7*(27), eabe9829. https://doi.org/10.1126/sciadv.abe9829

Conclusion

Humans and wildlife have coexisted over millennia but heightened human activities have disrupted this balance by placing greater demands on natural resources, resulting in negative impacts on wildlife (Chapin III et al., 2000; Storch et al., 2022). Amid the contemporary biodiversity crisis, effective management of the tensions between natural world and human communities is more critical than ever. Furthermore, not only do humans and nature compete for limited land resources, but different natural objectives (e.g. biodiversity, carbon, and water) also contend with each other (Bennett et al., 2009). As Chapter 2 has indicated, even with establishment of protected areas, connections between elephant populations in Tanzania are facing severe and potentially unmanageable pressure from human activities. Preserving large mammal species, such as African elephants, is particularly challenging since it demands vast territories to ensure viable populations (Tshipa et al., 2017). Achieving this often involves the implementation of large-scale strategies that extend beyond the boundaries of protected zones and interconnect nearby and distant ecosystems.

This dissertation aimed to examine how elephant populations are affected by a labyrinth of anthropological barriers that fragment once-extensive natural landscapes, and furthermore, to explore potential solutions for optimizing land allocation that can alleviate these impacts. While progressing through each chapter, I also explored the application of new Earth Observation (EO) datasets and techniques into conservation management. The subsequent section offers a concise summary of my research findings, followed by discussion into the implications they hold for biodiversity conservation in the years to come.

Advances in Earth Observation (EO) and Geospatial Artificial Intelligence (GeoAI) bring new opportunities as well as challenges for conservation studies

Until recently, land cover mapping was a major challenge in Africa's complex savanna landscapes, which exhibit substantial variations in vegetation cover over space and time (Solbrig, 1996). One possible reason is the technical challenge of separating the woody and herbaceous components in savanna biomes (Whitley et al., 2017). The land cover maps that are accessible for this region are usually part of global products (Buchhorn, Lesiv, et al., 2020; Buchhorn, Smets, et al., 2020; Congalton et al., 2017; Xu et al., 2019), with savanna being a land cover category that is often neglected due to its high uncertainty (Schmitt et al., 2020).

In my first chapter, I employed new Earth Observation (EO) data and GeoAI techniques (e.g. Deep Learning) to map Tanzania and revealed the opportunities and challenges that it presents for conservation studies. Upon examining various widely-used global land cover products (Buchhorn, Lesiv, et al., 2020; Buchhorn, Smets, et al., 2020; Congalton et al., 2017; Xu et al., 2019), I discovered that when using these datasets in African savanna environments at the regional scale, it is necessary to exercise extra caution due to the risk of high error rates, which may render downstream analyses ineffective (Estes et al., 2018). To address this issue in my analysis, I endeavored to create a new land cover map. To achieve this, I developed a deep learning model by combining high spatial resolution PlanetScope imagery with high temporal resolution Sentinel-1 imagery. Through this approach, I observed that an increase in spatial resolution significantly enhances object boundary delineation (e.g., for smallholder farms), improves detection of minor land cover types (e.g., residential), and describes landscape patches

more accurately (Abdi et al., 2022; Jin et al., 2019; Kerner et al., 2020). Furthermore, high temporal resolution enables a more comprehensive temporal signature of landscape objects, thereby enhancing the ability to differentiate between different land cover features. These advantages provide a promising solution to land cover issues in African savanna ecosystems and significantly benefit conservation studies.

While mapping, I also encountered several challenges, one of which was the absence of reliable reference labels for training the model (Ma et al., 2019; Yuan et al., 2020). Despite the growing availability of geospatial data via platforms such as Radiant Earth Foundation (https://radiant.earth), the volume and quality of data is still insufficient to train a fully developed deep learning model for not so well-attended places (Abdi et al., 2022). It has been a significant bottleneck for land cover mapping in savanna, despite the abundance of satellite images and advanced models at our disposal. This challenge motivated me to develop a method for quickly gathering extensive land cover labels. I argue that continuous efforts are required to fill these data gaps in the near future to push forward the application of advanced EO data and GeoAI techniques in savanna ecosystems. Additionally, from an open data perspective, I support establishing standards for data processing and product evaluation (Elmes et al., 2020), as well as making the raw data and model structures publicly open, to ensure the reproducibility and transferability of the models and data.

The resulting map made in Chapter 1 exhibited noticeable improvements in land cover type differentiation relative to existing land cover products, particularly in cropland and grassland. The study also suggests several possible avenues for future improvements. In African landscapes, fallow and abandoned croplands are common and critical land-use types that contribute to both food security and biodiversity (Crawford et al., 2022; Tong et al., 2020).

However, they are often excluded as a target class in many land cover studies. Future investigations should incorporate these land cover types into the mapping process (Tong et al., 2020). Second, in Chapter 1, a land cover map of 2018 was created, but time-series land cover maps are essential for understanding land changes and their impacts on biodiversity in a region. Therefore, creating land cover maps of the same high quality and resolution for multiple years and regularly updating them is necessary and an area that I intend to pursue in future work.

Elephants are confronted with unprecedented threats from human disturbances, but these can be mitigated with proper land management

After getting the land cover map, Chapter 2 & 3 analyzes the habitat suitability for elephants factoring in anthropological pressures, and analyzes the correlation between elephant occurrence and landscape components and features across multiple spatial scales. At broad scales, the findings show that human disturbances, such as settlement density, road density, and cropland edge density, have a more significant impact on elephant habitat selection than natural features. Thus, it is apparent that human activities have played a significant role in shaping the distribution of elephants in Tanzania. Moreover, I noticed that water resources are slightly more critical for shaping elephant distribution than food resources. Thus I argue that preserving the open connection to water resources is a critical broad-scale conservation strategy for elephants (Chamaillé-Jammes et al., 2007; Tshipa et al., 2017; Wato et al., 2018; Wood et al., 2022). Within broad regions of generally suitable and protected habitat, elephants' distribution and

movement at fine scales are primarily influenced by the availability of food resources, while still avoiding areas with high levels of human activity.

In suitability modeling, my multi-scale approach involved combining environmental features at both broad and fine scales, as well as utilizing fuzzy polygon-based and precise location-based survey observations of elephant distribution. In the process, I observed that combining environmental suitability obtained at coarse scales can improve prediction accuracy and certainty at fine scales. More importantly, combining multiple data sources that are considerably different in design and accuracy–for instance, the expert range map made by IUCN and occurrence survey dataset shared in GBIF–can overcome their respective shortcomings (e.g. sampling bias), thereby improving model reliability. Given that collecting data is still expensive and time-consuming, and certain areas remain inaccessible, I suggest that more efforts should be directed towards enhancing the capability and adaptability of integrating various species distribution data (Chevalier et al., 2021; Fletcher et al., 2019; Gilbert et al., 2021; Isaac et al., 2020; Koshkina et al., 2017). This would enable us to maximize the use of existing datasets to their full potential and reduce the need for future investment in data collection.

Having determined the spatial suitability, my next step was to evaluate the landscape connectivity using Circuitscape. In Chapter 3, I began by examining the long-distance connectivity between the three primary elephant habitat clusters, considering the landscape as a whole. I discovered that some long-distance corridors used by elephant populations have been entirely obstructed by humans. Thus, in Chapter 4, I conducted a more in-depth analysis to minimize the impact of future agricultural expansion on elephant movement. In Chapter 3, I conducted a similar connectivity analysis to evaluate connectivity within each major habitat cluster, which highlighted that several protected zones (> 8% of the whole areas) have been

encroached upon by agriculture. Although these croplands do not cover vast areas, they happen to obstruct many vital elephant movement corridors. Therefore, immediate conservation actions, such as community-based conservation strategies (Berkes, 2004; Brooks, 2016; Galvin et al., 2018), should be implemented to maintain necessary connectivity through the landscape that has been fragmented by human disturbances. Otherwise, the elephants in Tanzania will soon be isolated into disconnected habitat fragments.

Nonetheless, there are many other factors that should be considered in decision-making. Besides agriculture, mining is another human activity that has significant impacts on elephant conservation (Edwards et al., 2014; Rija et al., 2013; Seki et al., 2022). It can cause habitat destruction, fragmentation, and degradation of soil and water resources. Unfortunately, these impacts were not considered in this analysis. Future studies should include the impacts of mining and other human activities to comprehensively evaluate the threats to elephant conservation. Additionally, a modeling-only approach was employed to evaluate landscape connectivity, which may not reflect the actual conditions on the ground. Although evidence-based methods utilizing real-world tracking data are preferable, they require substantial investments in data collection and the establishment of guidelines for sharing sensitive conservation data.

Conservation requires forward-looking and multidimensional solutions

Finding a balance between local security and development concerns and international interests in conserving threatened species remains one of the biggest challenges in conservation today (Moss et al., 2011; Nyhus, 2016). Since the ban of ivory trade, elephant populations across Africa, particularly in Tanzania, have begun to recover (Thouless et al., 2016). However, this,

coupled with agricultural expansion, has resulted in an increase in the incidence of human-elephant conflicts (Dudley & Alexander, 2017; Haddad et al., 2015). To reduce these conflicts, a critical first step is to understand the spatial distribution of agriculture and the cost of land conversion on elephant conservation. Through this process, I discovered that critical corridors/hotspots exist outside of protected areas that should be protected. Compared to ecological indicators such as biodiversity, the importance of land units for preserving elephant habitat connectivity exhibits a weaker association with the distribution of protected areas. It may indicate that prior conservation strategies (e.g. establishment of protected areas) provide good protection for biodiversity, but may lack account for habitat connectivity.

In Chapter 4, I build a land-use prioritization model that can optimize areas for agricultural expansion with maximum production benefits and minimum ecological costs. Considering that other ecological goals, such as conserving biodiversity and carbon stocks are also important, I also incorporated the cost of land conversion on biodiversity and carbon into the analysis. By assigning varying weights to different ecological goals in the decision-making process, I discovered that tradeoffs between different ecological objectives are unavoidable. However, achieving a solution that balances different ecological demands may be more advantageous than maximizing one single objective. This suggests that conservation decision-making should consider multiple dimensions, even if there is currently only one main focus.

One critical factor in the land-use prioritization model is the assumption that crop yield will increase in the future due to agricultural intensification strategies. The extent to which the current yield gap can be closed is a crucial factor in land allocation decisions. With the full yield gap closed, the result suggests that Tanzania can accommodate agriculture to meet future food demand while also preserving biodiversity and reserving a sustainable amount of spare land.

Under a scenario in which half of the yield gap is closed, all the remaining non-agricultural areas would have to be converted to cropland to meet future food demand (4 times the current production). If major yield improvements are not achieved, Tanzania will either need to undertake a massive cropland expansion, or will be unable to produce the food requirements specified under this projection. This, in turn, could exacerbate the issue of human-wildlife conflicts.

While our study offers valuable insights on land-use prioritization by considering landscape connectivity, there are non-negligible limitations and uncertainties. The analysis is sensitive to errors in the input data, particularly the uncertainties in current and attainable yield estimation. These uncertainties can lead to ineffective and sometimes incorrect conservation decisions for a given region. Therefore, it is necessary to invest extra efforts to continuously improve the quality of data. Additionally, incorporating uncertainties within the analysis workflow can enhance the robustness of the results. Our analysis only examined a few crops (maize, pulses, rice, cassava), crops are not always grown where they can be grown. These variations can affect the analysis result significantly.

References

Abdi, A. M., Brandt, M., Abel, C., & Fensholt, R. (2022). Satellite Remote Sensing of Savannas: Current Status and Emerging Opportunities. *Journal of Remote Sensing*, *2022*, 1–20. https://doi.org/10.34133/2022/9835284

Battude, M., Al Bitar, A., Morin, D., Cros, J., Huc, M., Sicre, C. M., Le Dantec, V., & Demarez, V. (2016). Estimating maize biomass and yield over large areas using high spatial and temporal resolution Sentinel-2 like remote sensing data. *Remote Sensing of Environment*, *184*, 668–681.

Benediktsson, J. A., Chanussot, J., & Moon, W. M. (2012). Very high-resolution remote sensing: Challenges and opportunities [point of view]. *Proceedings of the IEEE*, *100*(6), 1907–1910.

Bennett, E. M., Peterson, G. D., & Gordon, L. J. (2009). Understanding relationships among multiple ecosystem services: Relationships among multiple ecosystem services. *Ecology Letters*, *12*(12), 1394–1404. https://doi.org/10.1111/j.1461-0248.2009.01387.x

Berkes, F. (2004). Rethinking Community-Based Conservation. *Conservation Biology*, *18*(3), 621–630. https://doi.org/10.1111/j.1523-1739.2004.00077.x

Brooks, J. (2016). Recognizing the many possible outcomes of community-based conservation. *The Routledge Handbook of Philosophy of Biodiversity*, 294.

Buchhorn, M., Lesiv, M., Tsendbazar, N.-E., Herold, M., Bertels, L., & Smets, B. (2020). Copernicus global land cover layers—Collection 2. *Remote Sensing*, *12*(6), 1044.

Buchhorn, M., Smets, B., Bertels, L., De Roo, B., Lesiv, M., Tsendbazar, N.-E., Herold, M., & Fritz, S. (2020). Copernicus Global Land Service: Land Cover 100m: Collection 3 Epoch 2018, Globe. *Version V3. 0.1)[Data Set]*.

Chamaillé-Jammes, S., Valeix, M., & Fritz, H. (2007). Managing heterogeneity in elephant distribution: Interactions between elephant population density and surface-water availability: Surface water and elephant distribution. *Journal of Applied Ecology*, *44*(3), 625–633. https://doi.org/10.1111/j.1365-2664.2007.01300.x

Chapin III, F. S., Zavaleta, E. S., Eviner, V. T., Naylor, R. L., Vitousek, P. M., Reynolds, H. L., Hooper, D. U., Lavorel, S., Sala, O. E., Hobbie, S. E., Mack, M. C., & Díaz, S. (2000). Consequences of changing biodiversity. *Nature*, *405*(6783), 234–242. https://doi.org/10.1038/35012241

Chevalier, M., Broennimann, O., Cornuault, J., & Guisan, A. (2021). Data integration methods to account for spatial niche truncation effects in regional projections of species distribution. *Ecological Applications*, *31*(7). https://doi.org/10.1002/eap.2427

Congalton, R., Yadav, K., McDonnell, K., Poehnelt, J., Stevens, B., Gumma, M., Teluguntla, P., & Thenkabail, P. (2017). *Global Food Security-support Analysis Data (GFSAD) Cropland Extent 2015 Validation 30 m V001*.

Crawford, C. L., Yin, H., Radeloff, V. C., & Wilcove, D. S. (2022). Rural land abandonment is too ephemeral to provide major benefits for biodiversity and climate. *Science Advances*, *8*(21), eabm8999. https://doi.org/10.1126/sciadv.abm8999

Dudley, N., & Alexander, S. (2017). Agriculture and biodiversity: A review. *Biodiversity*, *18*(2–3), 45–49. https://doi.org/10.1080/14888386.2017.1351892

Edwards, D. P., Sloan, S., Weng, L., Dirks, P., Sayer, J., & Laurance, W. F. (2014). Mining and the African environment. *Conservation Letters*, *7*(3), 302–311.

Elmes, A., Alemohammad, H., Avery, R., Caylor, K., Eastman, J., Fishgold, L., Friedl, M., Jain, M., Kohli, D., Laso Bayas, J., Lunga, D., McCarty, J., Pontius, R., Reinmann, A., Rogan, J., Song, L., Stoynova, H., Ye, S., Yi, Z.-F., & Estes, L. (2020). Accounting for Training Data Error in Machine Learning Applied to Earth Observations. *Remote Sensing*, *12*(6), 1034. https://doi.org/10.3390/rs12061034

Estes, L., Chen, P., Debats, S., Evans, T., Ferreira, S., Kuemmerle, T., Ragazzo, G., Sheffield, J., Wolf, A., Wood, E., & Caylor, K. (2018). A large-area, spatially continuous assessment of land cover map error and its impact on downstream analyses. *Global Change Biology*, *24*(1), 322–337. https://doi.org/10.1111/gcb.13904

Fletcher, R. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A., & Dorazio, R. M. (2019). A practical guide for combining data to model species distributions. *Ecology*, e02710. https://doi.org/10.1002/ecy.2710

Galvin, K. A., Beeton, T. A., & Luizza, M. W. (2018). African community-based conservation: A systematic review of social and ecological outcomes. *Ecology and Society*, *23*(3). https://doi.org/10.5751/ES-10217-230339

Gilbert, N. A., Pease, B. S., Anhalt-Depies, C. M., Clare, J. D. J., Stenglein, J. L., Townsend, P. A., Van Deelen, T. R., & Zuckerberg, B. (2021). Integrating harvest and camera trap data in species distribution models. *Biological Conservation*, *258*, 109147. https://doi.org/10.1016/j.biocon.2021.109147

Haddad, N. M., Brudvig, L. A., Clobert, J., Davies, K. F., Gonzalez, A., Holt, R. D., Lovejoy, T. E., Sexton, J. O., Austin, M. P., Collins, C. D., Cook, W. M., Damschen, E. I., Ewers, R. M., Foster, B. L., Jenkins, C. N., King, A. J., Laurance, W. F., Levey, D. J., Margules, C. R., … Townshend, J. R. (2015). Habitat fragmentation and its lasting impact on Earth's ecosystems. *Science Advances*, *1*(2), e1500052. https://doi.org/10.1126/sciadv.1500052

Isaac, N. J. B., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., Freeman, S. N., Golding, N., Guillera-Arroita, G., Henrys, P. A., Jarvis, S., Lahoz-Monfort, J., Pagel, J., Pescott, O. L., Schmucki, R., Simmonds, E. G., & O'Hara, R. B. (2020). Data Integration for Large-Scale Models of Species Distributions. *Trends in Ecology & Evolution*, *35*(1), 56–67. https://doi.org/10.1016/j.tree.2019.08.006

Jin, Z., Azzari, G., You, C., Di Tommaso, S., Aston, S., Burke, M., & Lobell, D. B. (2019). Smallholder maize area and yield mapping at national scales with Google Earth Engine. *Remote Sensing of Environment*, *228*, 115–128. https://doi.org/10.1016/j.rse.2019.04.016

Kerner, H., Tseng, G., Becker-Reshef, I., Nakalembe, C., Barker, B., Munshell, B., Paliyam, M., & Hosseini, M. (2020). Rapid Response Crop Maps in Data Sparse Regions. *ArXiv:2006.16866 [Cs, Eess]*. http://arxiv.org/abs/2006.16866

Koshkina, V., Wang, Y., Gordon, A., Dorazio, R. M., White, M., & Stone, L. (2017). Integrated species distribution models: Combining presence-background data and site-occupancy data with imperfect detection. *Methods in Ecology and Evolution*, *8*(4), 420–430. https://doi.org/10.1111/2041-210X.12738

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *152*, 166–177. https://doi.org/10.1016/j.isprsjprs.2019.04.015

Moss, C. J., Croze, H., & Lee, P. C. (2011). *The Amboseli elephants: A long-term perspective on a long-lived mammal*. University of Chicago Press.

Nyhus, P. J. (2016). Human–Wildlife Conflict and Coexistence. *Annual Review of Environment and Resources*, *41*(1), 143–171. https://doi.org/10.1146/annurev-environ-110615-085634

Rija, A., Kideghesho, J., Mwamende, K. A., & Selemani, I. (2013). *Emerging issues and challenges in conservation of biodiversity in the rangelands of Tanzania*.

Schmitt, M., Prexl, J., Ebel, P., Liebel, L., & Zhu, X. X. (2020). *Weakly Supervised Semantic Segmentation of Satellite Images for Land Cover Mapping—Challenges and Opportunities* (arXiv:2002.08254). arXiv. http://arxiv.org/abs/2002.08254

Seki, H. A., Thorn, J. P., Platts, P. J., Shirima, D. D., Marchant, R. A., Abeid, Y., Baker, N., Annandale, M., & Marshall, A. R. (2022). Indirect impacts of commercial gold mining on adjacent ecosystems. *Biological Conservation*, *275*, 109782.

Solbrig, O. T. (1996). The diversity of the savanna ecosystem. In *Biodiversity and savanna ecosystem processes* (pp. 1–27). Springer.

Storch, D., Šímová, I., Smyčka, J., Bohdalková, E., Toszogyova, A., & Okie, J. G. (2022). Biodiversity dynamics in the Anthropocene: How human activities change equilibria of species richness. *Ecography*, *2022*(4), ecog.05778. https://doi.org/10.1111/ecog.05778

Thouless, C., Dublin, H. T., Blanc, J., Skinner, D., Daniel, T., Taylor, R., Maisels, F., Frederick, H., & Bouché, P. (2016). African elephant status report 2016. *Occasional Paper Series of the IUCN Species Survival Commission*, *60*.

Tong, X., Brandt, M., Hiernaux, P., Herrmann, S., Rasmussen, L. V., Rasmussen, K., Tian, F., Tagesson, T., Zhang, W., & Fensholt, R. (2020). The forgotten land use class: Mapping of fallow fields across the Sahel using Sentinel-2. *Remote Sensing of Environment*, *239*, 111598. https://doi.org/10.1016/j.rse.2019.111598

Tshipa, A., Valls-Fox, H., Fritz, H., Collins, K., Sebele, L., Mundy, P., & Chamaillé-Jammes, S. (2017). Partial migration links local surface-water management to large-scale elephant conservation in the world's largest transfrontier conservation area. *Biological Conservation*, *215*, 46–50. https://doi.org/10.1016/j.biocon.2017.09.003

Wato, Y. A., Prins, H. H., Heitkönig, I. M., Wahungu, G. M., Ngene, S. M., Njumbi, S., & Van Langevelde, F. (2018). Movement patterns of African elephants (Loxodonta africana) in a semi-arid savanna suggest that they have information on the location of dispersed water sources. *Frontiers in Ecology and Evolution*, *6*, 167.

Whitley, R., Beringer, J., Hutley, L. B., Abramowitz, G., De Kauwe, M. G., Evans, B., Haverd, V., Li, L., Moore, C., Ryu, Y., & others. (2017). Challenges and opportunities in land surface modelling of savanna ecosystems. *Biogeosciences*, *14*(20), 4711–4732.

Wood, M., Chamaillé-Jammes, S., Hammerbacher, A., & Shrader, A. M. (2022). African elephants can detect water from natural and artificial sources via olfactory cues. *Animal Cognition*, *25*(1), 53–61.

Xu, Y., Yu, L., Feng, D., Peng, D., Li, C., Huang, X., Lu, H., & Gong, P. (2019). Comparisons of three recent moderate resolution African land cover datasets: CGLS-LC100, ESA-S2-LC20, and FROM-GLC-Africa30. *International Journal of Remote Sensing*, *40*(16), 6185–6202.

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., & others. (2020). Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing of Environment*, *241*, 111716.

Appendix A
Appendix to Chapter 1

Published with:

Song, L., Estes, A. B., & Estes, L. D. (2023). A super-ensemble approach to map land cover types with high resolution over data-sparse African savanna landscapes. *International Journal of Applied Earth Observation and Geoinformation*, 116, 103152.

A.1 PlanetScope NICFI basemap

Planet Labs, Inc is a commercial satellite operator to harness the enhanced observing capability provided by standardized small satellites (CubeSats) (Planet Team, 2017). The primary PlanetScope (PS) constellation consists of over 200 CubeSats in a sun-synchronous orbit, which together provide near-daily imaging at 3-4 m resolution over four channels, three visible and one near-infrared (455–515 nm; 500–590 nm; 590–670 nm; 780–860 nm, respectively). The level-3B PS Analytic Ortho Scene Product is calibrated to surface reflectance and orthorectified with < 10 m RMSE geolocation accuracy (Marta, 2018).

The high spatial and temporal resolution of PS makes it particularly valuable for mapping landcover in rapidly changing, hard-to-map environments (Johansen et al., 2022), although the cost of these commercial data makes it difficult to apply in large area mapping studies. A recent collaboration between Norway's International Climate and Forests Initiative Imagery Program (NICFI), Planet, Airbus, and Kongsberg Satellite Service released the free of

charge and analysis-ready high-resolution (4.77 m per pixel) basemap quads of global tropic regions to facilitate better understanding and prevention of tropical deforestation. The broad coverage and free availability of this dataset will revolutionize land surface monitoring in these tropical countries.

The NICFI program provides historical biannual, analysis-ready basemaps covering the period December 2015 to August 2020, and near real-time monthly data since September 2020. The examples shown in Figure 1-2 indicate that all land cover types are more distinguishable from each other in season 1, within only water and forest/dense trees having clearly distinct signatures across both seasons. Bareland is distinct in season 2 and built-up in season 1.

A.2 Sentinel-1 Synthetic Aperture Radar (SAR)

The Sentinel-1 (S1) constellation carries a C-band SAR to provide a day-and-night supply of imagery at 10-20m resolution every 6 days. S1 imagery is effective for detecting agricultural land cover and is important for monitoring humid regions because it is not affected by cloud cover (Torres et al., 2012). In this study, we utilized level-1 Ground Range Detected (GRD) Interferometric Wide Swath (IW) images acquired with dual polarization (VV + VH), which were multi-looked and projected to ground range in square pixels with reduced speckle. To standardize the imagery, we applied a generic workflow proposed by Filipponi (2019) to preprocess GRD level-1 imagery.

After finishing all steps, the processed SAR images still contain significant speckling that can render them less effective for detecting fine-grained landscape features (e.g., smallholder farms). To reduce this speckle, we applied a guided filter (He et al., 2013; He & Sun,

2015), a smoothing technique that is designed to preserve edge features. We then applied a harmonic regression (Eq. (A-1) to both the VV and VH polarizations. Harmonic regression coefficients summarize critical temporal features, enhancing the ability to differentiate land cover types based on seasonal information contained in the series (Moody & Johnson, 2001), while significantly reduce the number of raw images to use.

$$f(t) = a + bt + \sum_{k=1}^{n} (c_k \sin\left(\frac{2\pi kt}{d_{yr}}\right) + d_k \cos\left(\frac{2\pi kt}{d_{yr}}\right)) \quad \text{(A-1)}$$

Where a, b, c_k, d_k, are intercept, slope, sine, and cosine coefficients. t is the time domain of the observations (2017-10-01 – 2018-09-30), d_{yr} is the overall day number of a year, and n is the number of harmonic pairs. We used n equal to 2 based on the seasonal features over the study area (Moody & Johnson, 2001). We applied the least absolute shrinkage and selection operator (Lasso) to estimate the robust coefficients of harmonic regression. The 6 bands of the generated image in order are intercept, slope, and coefficients of $\sin\left(\frac{2\pi t}{d_{yr}}\right)$, $\cos\left(\frac{2\pi t}{d_{yr}}\right)$, $\sin\left(\frac{2\pi \cdot 2t}{d_{yr}}\right)$, and $\cos\left(\frac{2\pi \cdot 2t}{d_{yr}}\right)$, which were processed into tiles using the NICFI quad grid.

A.3 Initial selection of Land cover (LC) products

We selected five land cover products for pre-assessment. They are ESA CCI land cover S2 prototype land cover 20m map of Africa 2016 (CCI-LC, 2020), GLOBELAND30 (2010) (Chen et al., 2017), ESRI Sentinel-2 10-Meter Land Use/Land Cover, Copernicus global land cover map (CGLS_LC) (Buchhorn et al., 2020), and Finer Resolution Observation and Monitoring – Global Land Cover (FROM-GLC) (Gong et al., 2019).

However, because the objective of making consensus is to get useful training labels, not all of the products were used to make labels due to the high disagreement between them. If using all of them, the area of the resultant census would be small and highly biased to the pixels that are the easy ones for any model to learn. It would harm the model generalization for both Random Forest and U-Net. ESA CCI S2 prototype land cover map underestimates cropland, overestimates shrubland over cropland area but underestimates it over other regions (Figure A-1). GLOBELAND30 extremely underestimates shrubland in our study area, even though it performs well on cropland (Figure A-1). Cropland is highly underestimated in the Esri Land Cover map (Figure A-1), which makes it less useful for our study area. Moreover, it only includes class rangeland. FROM-GLC also underestimates cropland, but it performs well on other land cover types. Copernicus global land cover map performs well on all classes. After the pre-assessment, Copernicus global land cover map (CGLS_LC) (Buchhorn et al., 2020), and Finer Resolution Observation and Monitoring – Global Land Cover (FROM-GLC) (Gong et al., 2019) were selected to make a consensus. The selected products were converted to the land cover types defined in this study following Table A-1.

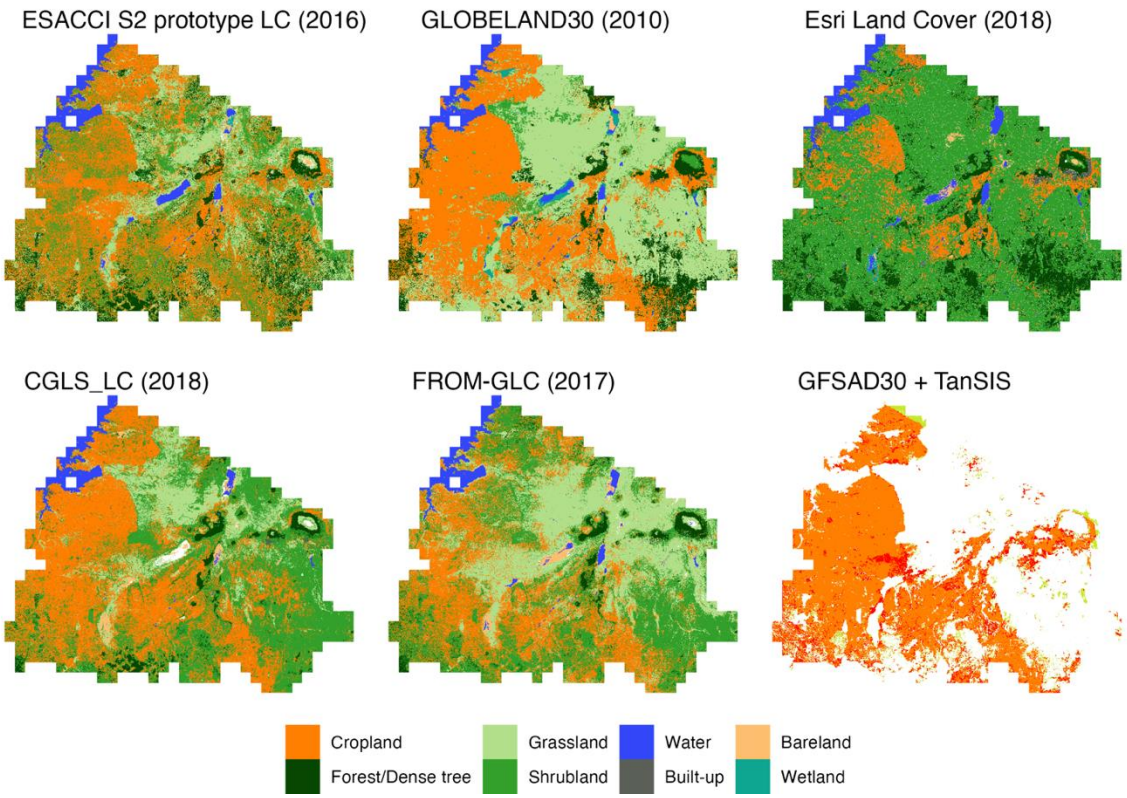

Figure A-1. Selected Land Cover products we evaluated in pre-assessment (Note: Esri Land Cover only has rangeland, which is colored the same as shrubland in the other two maps. In GFSAD +TanSIS map, orange is their overlap area, red is the area covered by TanSIS only, and yellow is the area covered by GFSAD30 only)

Table A-1. The conversion table from original classes to classes defined in this study

ID	Land cover	CGLS_LC100M	FROM-GLC 2017v1
1	Cropland	40	1
2	Forest/Dense tree	111, 112, 113, 114, 115, 116	2
3	Grassland	30	3
4	Shrubland	20, 121, 122, 123, 124, 125, 126	4
5	Water	80, 200	6
6	Bareland	60	9
7	Built-up	50	8
8	Wetland	90	5

The detailed evaluation of the selected land cover products is shown in Table A-2.

Table A-2. Evaluation of different LC products using the independent test dataset

Product	Class	TNR	NPV	UA	PA	BA	F1 score
CGLS_LC100M	Cropland	81.29%	86.73%	67.64%	75.87%	78.58%	71.52%
	Forest/Dense tree	98.69%	98.22%	70.31%	63.38%	81.04%	66.67%
	Grassland	93.38%	90.39%	60.38%	50.39%	71.89%	54.94%
	Shrubland	79.81%	83.42%	67.00%	72.10%	75.96%	69.46%
	Water	100.00%	99.73%	100.00%	91.67%	95.83%	95.65%
	Built-up	99.93%	97.40%	95.24%	33.90%	66.91%	50.00%
	Bareland	99.73%	98.81%	42.86%	14.29%	57.01%	21.43%
	Average	93.26%	93.53%	71.92%	57.37%	75.32%	61.38%
FROM-GLC 2017v1	Cropland	89.04%	80.51%	72.51%	57.30%	73.17%	64.02%
	Forest/Dense tree	98.42%	97.74%	60.87%	51.85%	75.14%	56.00%
	Grassland	78.03%	92.54%	37.05%	67.27%	72.65%	47.79%
	Shrubland	80.60%	78.35%	65.88%	62.71%	71.66%	64.26%
	Water	99.78%	100.00%	94.51%	100.00%	99.89%	97.18%
	Built-up	99.83%	96.76%	73.33%	12.50%	56.16%	21.36%
	Bareland	99.87%	99.08%	62.50%	18.52%	59.20%	28.57%
	Average	92.22%	92.14%	66.67%	52.88%	72.55%	54.17%
GFSAD30	Cropland	83.91%	73.75%	89.73%	82.48%	83.20%	85.96%
	Water	83.58%	90.71%	71.85%	83.04%	83.31%	77.04%
	Others	99.78%	99.65%	93.51%	90.00%	94.89%	91.72%
	Average	89.09%	88.04%	85.03%	85.17%	87.13%	84.91%
	TanSIS	78.66%	93.95%	67.75%	89.85%	84.26%	77.25%

A.4 U-Net structure and the relevant computation

The diagram of the U-Net architecture used in this study is shown in Figure A-2. The structure on the left side is the encoding path and the structure on the right side is the decoding. Every block in the encoding section has two convolutional layers and a max pooling layer. Every block in the decoding section uses bilinear up-sampling followed by several convolutional layers to produce dense features. The skip connections help to produce fine-grained segmentation

results, thereby improving U-Net's ability to segment an image. The input dimension in this study is $512 \times 512 \times 14$. Two outputs are obtained from the U-Net by using softmax: the class probabilities and land cover classification map.

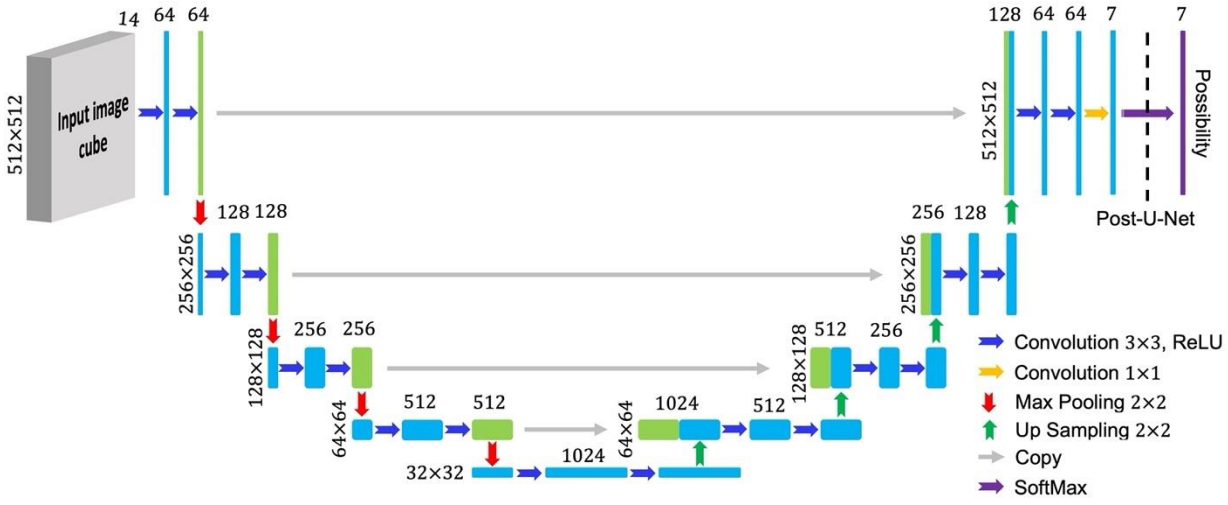

Figure A-2. Illustration of the U-Net structure for semantic segmentation (modified from (Ronneberger et al., 2015))

The gap-filled Random Forest model was trained and applied on a macOS with a processor of 2.3 GHz Quad-Core Intel Core i7 and Memory of 32 GB 3733 MHz LPDDR4X. It took 1-2 minutes using one core for Random Forest to predict one sub-tile. Final U-Net training and prediction were made with 32 batch size on AWS instance g4dn.12xlarge with 4 GPUs and 64 GB GPU memory. The price was $1.1819/hr for the SPOT instance when we queried. It spent ~13 hours to train the model for 200 epochs. It took ~1-2 hours to make the prediction.

A.5 Accuracy assessment metrics

User's accuracy describes the proportion of pixels mapped as a certain class that has the same reference class. The producer's accuracy is the proportion of pixels with a certain reference class that is mapped as the same class (Story & Congalton, 1986). The true negative rate

represents the proportion of pixels that do not have the reference of a certain class that is not mapped as this class either. The Negative predictive value means the proportion of pixels not mapped as a certain class that does not have the reference of this class (Barsi et al., 2018; Brodersen et al., 2010). The balanced accuracy is the average of the user's accuracy and producer's accuracy, and the F1 score can be interpreted as their weighted average (Brodersen et al., 2010).

Let p_{ii} denotes the number of pixels that are predicted to be class i and have reference class i,, and $-p_{ii}$ denotes the number of pixels that are not predicted to be class i and have other reference class except i. Then $p_{i.}$ indicates the number of pixels predicted to belong to class i, and $p_{.i}$ indicates the number of pixels that have reference of class i. Correspondingly, $-p_{i.}$ is the number of pixels predicted not to belong to class i, and $-p_{.i}$ indicates the number of pixels that have other reference classes except i. The metrics are calculated as follows (Olofsson et al., 2014):

$$TNR = \frac{-p_{ii}}{-p_{.i}} \tag{A-2}$$

$$NPV = \frac{-p_{ii}}{-p_{i.}} \tag{A-3}$$

$$UA = \frac{p_{ii}}{p_{i.}} \tag{A-4}$$

$$PA = \frac{p_{ii}}{p_{.i}} \tag{A-5}$$

$$BA = \frac{TNR + PA}{2} \tag{A-6}$$

$$F1\ score = 2 \times \frac{UA \times PA}{UA + PA} \tag{A-7}$$

For U-Net training, we also used intersection over union (IoU) and its average (mIoU) as an evaluation metric that is calculated as follows:

$$IoU = \frac{p_{ii}}{p_{i.} \cup p_{.i}} \tag{A-8}$$

$$mIoU = \frac{1}{k}\sum_{i=1}^{k} IoU_i \tag{A-9}$$

A.6 Analysis of weight parameters of quality-weighted and class-balanced loss

As introduced in section 1.3.3.3, a quality-weighted loss function was implemented in our study to increase the model's resistance to noises. The growth rate k of the logistic function, used to calculate the correctness weight (w_c) and s_{max}, used to calculate the difficulty weight (w_d) were the two main parameters to tune. We did a simple sensitivity analysis on these two parameters to investigate their influence on model performance. We first set $s_{max} = 1$ to experiment with k, then used the best k to experiment with s_{max}. In each experiment, we trained a model for 30 epochs and calculated the mean and standard deviation of the average accuracy (AA) for the last 5 epochs to evaluate performance sensitivity (Figure A-3).

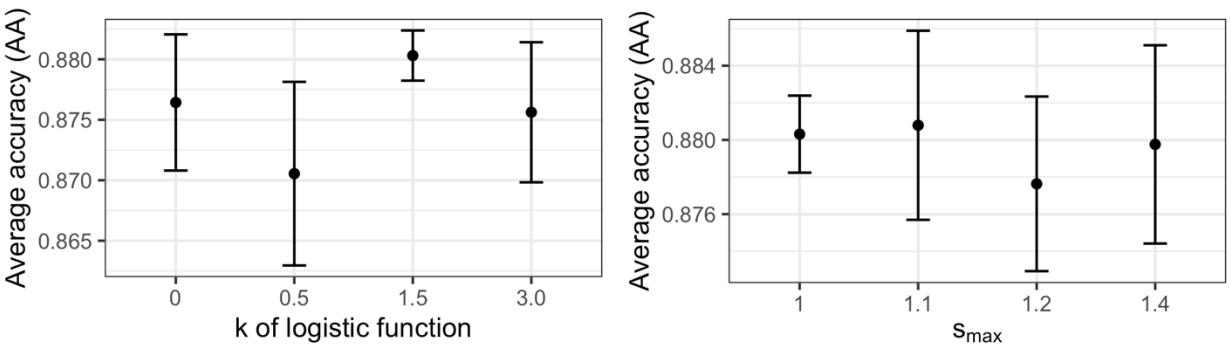

Figure A-3. Sensitivity analysis on weight parameters of quality-weighted and class-balanced loss

Parameter k varied from 0 to 3.0 with an increasing interval of 0.5, 1.0, and 1.5, with 0 indicating equal weights for all labels, and larger values indicating increasingly lower weights for labels with low correctness. A too-large value is not recommended for k because the logistic curve would become too deep and would therefore discount many labels with moderate quality. In our case, when k is set to 1.5, the best performance (AA = 0.880 ± 0.002) was achieved.

Parameter s_{max} varied from 1 to 1.4 with intervals 0.1, 0.1, and 0.2, with $[1, s_{max}]$ being the scale of difficulty weights used in loss calculation. Here 1.4 was used as a large value for the sake of comparison. Since difficulty weights act as a model generalization, an s_{max} value that is too high might cause the model to overfit on difficult labels, therefore it is better to use lower s_{max} values. According to Figure A-3, values from 1 to 1.1 were reasonable choices for this study.

A.7 Other supplementary tables and figures

Table A-3. Estimated time spent by one interpreter on label creation, comparing time required for a hypothetical case in which tiles were labelled manually based on visual image interpretation to the estimated time spent editing labels using the semi-automated labelling approach developed in this study.

Label creation method	**Category**	**No of labels**	**Average label creation/editing time per tile**	**Total time**
Visual interpretation	Tile manually digitized	2572	60 minutes	154,320 minutes
Landcover consensus and Random Forest gap-filling	No edits	1013	0 min	0 minutes
	Light edits (Easy visual interpretation, set parameters for script to fix the label)	1322	1 min	1322 minutes
	Heavy edits (Hard visual interpretation and intensive hand drawing)	237	3 mins	711 minutes
	Overall	2572		2033 minutes

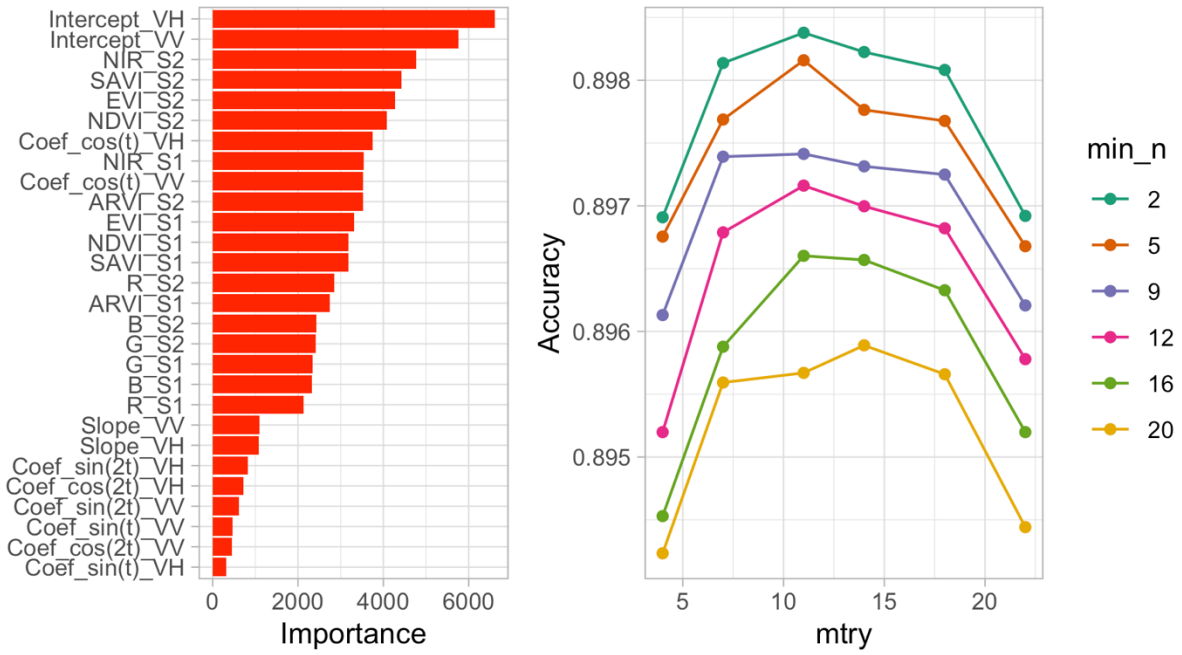

Figure A-4. Variable importance and tuning curve of random forest guessing model (S1 stands for NICFI season 1 image, S2 stands for NICFI season 2 images. B, G, R, and NIR are four spectral bands and NDVI, EVI, SAVI, and ARVI are four vegetarian indices. VV and VH are two types of polarization of Sentinel-1 images. Intercept, slope, and coefs are coefficients of harmonic regression fitting. sin/cos(t) represents one intra-annual seasonal cycle and sin/cos(2t) describes two intra-annual cycles. See more details in Section 1.3.2.1, A.1 and A.2)

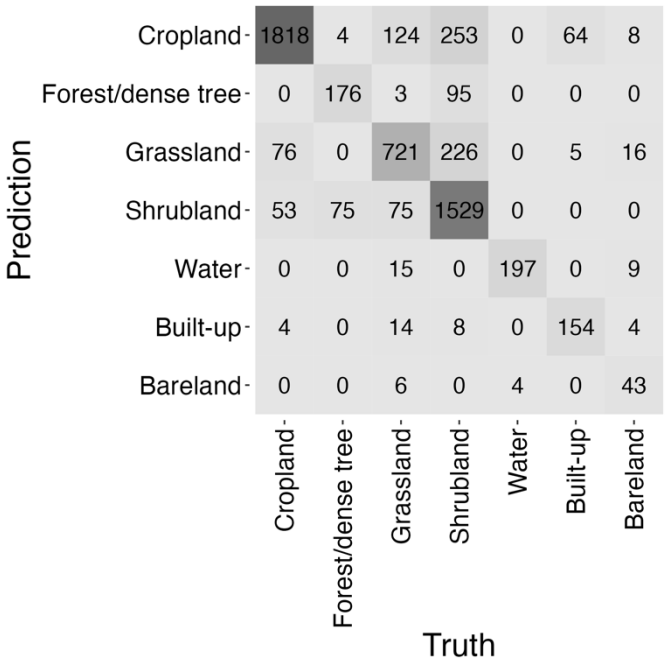

Figure A-5. Confusion matrix heatmap of the independent test of Random Forests model

References

Barsi, Á., Kugler, Zs., László, I., Szabó, Gy., & Abdulmutalib, H. M. (2018). Accuracy Dimensions in Remote Sensing. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLII–3*, 61–67. https://doi.org/10.5194/isprs-archives-XLII-3-61-2018

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. *2010 20th International Conference on Pattern Recognition*, 3121–3124.

Buchhorn, M., Smets, B., Bertels, L., De Roo, B., Lesiv, M., Tsendbazar, N.-E., Herold, M., & Fritz, S. (2020). Copernicus Global Land Service: Land Cover 100m: Collection 3 Epoch 2018, Globe. *Version V3. 0.1)[Data Set]*.

CCI-LC, E. (2020). *S2 Prototype Land cover 20m map of Africa 2016. Land Cover project of the ESA Climate Change Initiative*.

Chen, J., Cao, X., Peng, S., & Ren, H. (2017). Analysis and applications of GlobeLand30: A review. *ISPRS International Journal of Geo-Information*, *6*(8), 230.

Filipponi, F. (2019). Sentinel-1 GRD Preprocessing Workflow. *Proceedings*, *18*(1), 11. https://doi.org/10.3390/ECRS-3-06201

Gong, P., Liu, H., Zhang, M., Li, C., Wang, J., Huang, H., Clinton, N., Ji, L., Li, W., Bai, Y., Chen, B., Xu, B., Zhu, Z., Yuan, C., Ping Suen, H., Guo, J., Xu, N., Li, W., Zhao, Y., … Song, L. (2019). Stable classification with limited sample: Transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017. *Science Bulletin*, *64*(6), 370–373. https://doi.org/10.1016/j.scib.2019.03.002

He, K., & Sun, J. (2015). Fast Guided Filter. *ArXiv:1505.00996 [Cs]*. http://arxiv.org/abs/1505.00996

He, K., Sun, J., & Tang, X. (2013). Guided Image Filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(6), 1397–1409. https://doi.org/10.1109/TPAMI.2012.213

Johansen, K., Ziliani, M. G., Houborg, R., Franz, T. E., & McCabe, M. F. (2022). CubeSat constellations provide enhanced crop phenology and digital agricultural insights using daily leaf area index retrievals. *Scientific Reports*, *12*(1), 5244. https://doi.org/10.1038/s41598-022-09376-6

Marta, S. (2018). Planet Imagery Product Specifications. *Planet Labs: San Francisco, CA, USA*, 91.

Moody, A., & Johnson, D. M. (2001). Land-Surface Phenologies from AVHRR Using the Discrete Fourier Transform. *Remote Sensing of Environment*, *75*(3), 305–323. https://doi.org/10.1016/S0034-4257(00)00175-9

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., & Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, *148*, 42–57. https://doi.org/10.1016/j.rse.2014.02.015

Planet Team. (2017). Planet application program interface: In space for life on Earth. *San Francisco, CA*, *2017*, 40.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241.

Story, M., & Congalton, R. G. (1986). Accuracy assessment: A user's perspective. *Photogrammetric Engineering and Remote Sensing*, *52*(3), 397–399.

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., & others. (2012). GMES Sentinel-1 mission. *Remote Sensing of Environment*, *120*, 9–24.

Appendix B

Appendix to Chapter 2

Published with:

Song, L., & Estes, L. (2023). *itsdm*: Isolation forest-based presence-only species distribution modelling and explanation in R. *Methods in Ecology and Evolution*.

B.1 Supplementary tables, figures, and scripts

B.1.1 Supplementary tables and figures

Table B-1. Decisive arguments setting to specify the model type.

Model type	**Decisive arguments setting in isotree_po (Cortes, 2022)**
Isolation Forest (iForest)	ndim = 1
Extended isolation forest (EIF)	ndim > 1
Non-random split[1]	
Average gain	prob_pick_avg_gain > 0
	prob_split_avg_gain > 0 (only for ndim = 1)
Pooled gain	prob_pick_pooled_gain > 0
	prob_split_pooled_gain > 0 (only for ndim = 1)

[1] Non-random splitting benefits model accuracy but with the cost of less generalizable models. Regarding habitat modeling, non-random splitting is more suitable for mapping rare species or species with narrow environmental conditions. Cortes (2022) described that using a sub-sampling strategy and a deeper tree depth might benefit non-random split models.

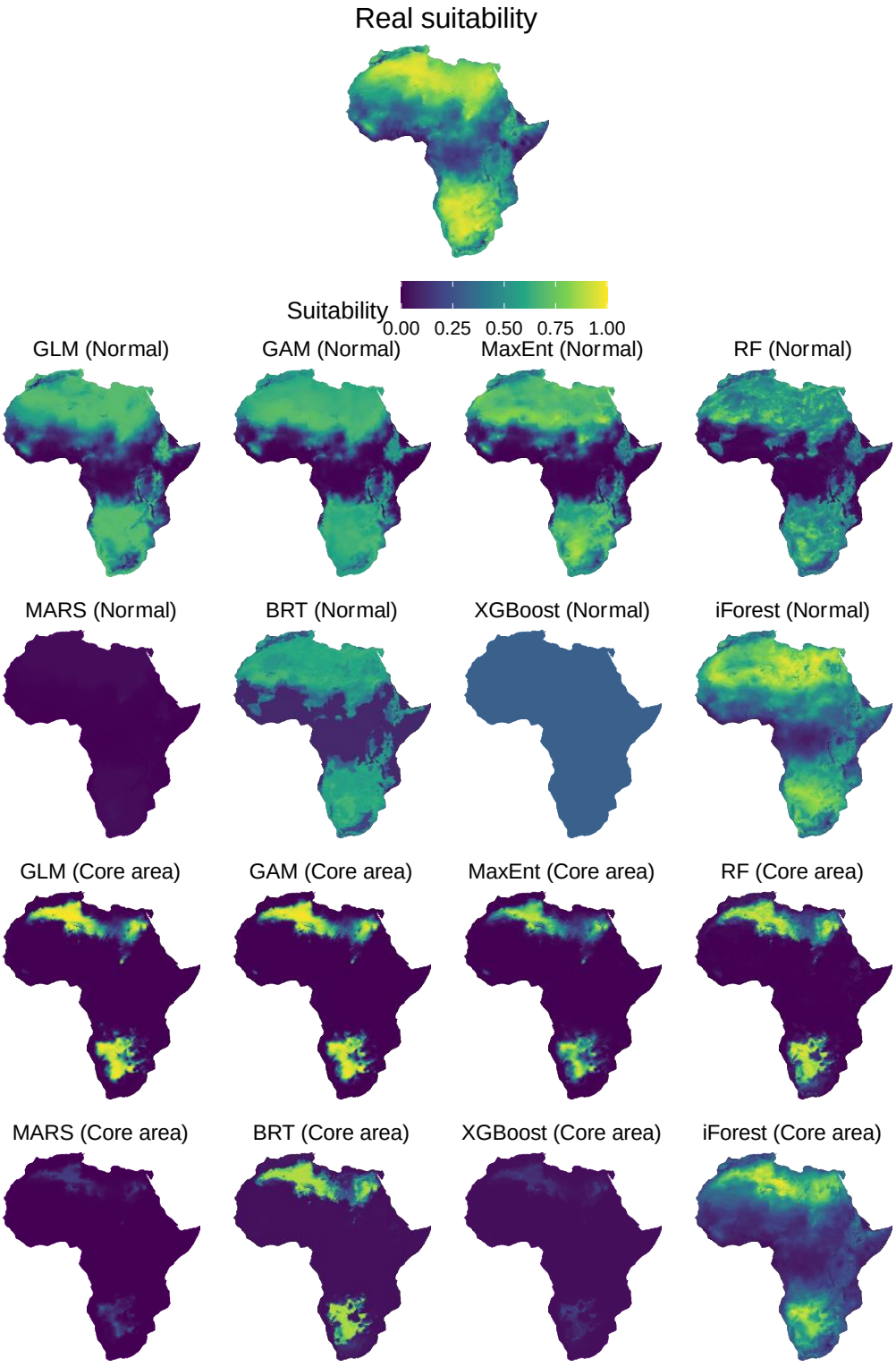

Figure B-1.Environmental suitability maps of No.6 virtual species predicted by different SDMs using different types of samples

B.1.2 Full version of the example

B.1.2.1 Generating a virtual species

Using a virtual species is a classic way to examine a model in ecological modelling. In this paper, we thus generate a virtual species to introduce and discuss how *itsdm* works. Our virtual species responds to climatic variables BIO1 (Annual Mean Temperature), BIO5 (Max Temperature of Warmest Month), and BIO12 (Annual Precipitation) in mainland Africa (Figure B-2). Presence-absence map is converted from the suitability map with a logistic function and a beta of 0.7. Then we extract 2000 presence-only samples from the presence-absence map as a virtual occurrence dataset to use.

```
# Load libraries
library(itsdm)
library(dplyr)
library(stars)
library(virtualspecies)
library(ggplot2)
library(ggplotify)
library(ggpubr)
select <- dplyr::select

# Prepare environmental variables
data("mainland_africa")
bios <- worldclim2(
  var = 'bio',
  bry = mainland_africa,
  path = tempdir(),
  nm_mark = 'africa') %>%
  st_normalize()

bios_sub <- bios %>% slice('band', c(1, 5, 12))
bios_sub <- as(bios_sub, 'Raster')

# Formatting of the response functions
set.seed(10)
my.parameters <- formatFunctions(
  bio1 = c(fun = 'dnorm', mean = 25, sd = 5),
  bio5 = c(fun = 'dnorm', mean = 35, sd = 5),
```

```
  bio12 = c(fun = 'dnorm', mean = 1000, sd = 500))

# Generation of the virtual species
my.species <- generateSpFromFun(
  raster.stack = bios_sub,
  parameters = my.parameters)

# Conversion to presence-absence
my.species <- convertToPA(
  my.species,
  beta = 0.7,
  plot = FALSE)

# Make pseudo presence-only samples
# Sampling
set.seed(12)
po.points <- sampleOccurrences(
  my.species,
  n = 2000,
  type = "presence only",
  plot = FALSE)
po_df <- po.points$sample.points %>%
  dplyr::select(x, y)

# Clean up
rm(po.points, my.parameters, bios_sub)

# Check the suitability
g_suit <- ggplot() +
  geom_stars(
    data = st_as_stars(my.species$suitab.raster),
    na.action = na.omit) +
  scale_fill_viridis_c("Suitability") +
  coord_equal() +
  theme_void() +
  theme(text = element_text(size = 10),
        legend.position = "bottom",
        plot.title = element_text(hjust = 0.5),
        legend.key.height = unit(0.4,"cm"))

# Check response curves
g_rsp <- as.ggplot(~plotResponse(my.species))

ggarrange(g_suit, g_rsp, ncol = 2)
```

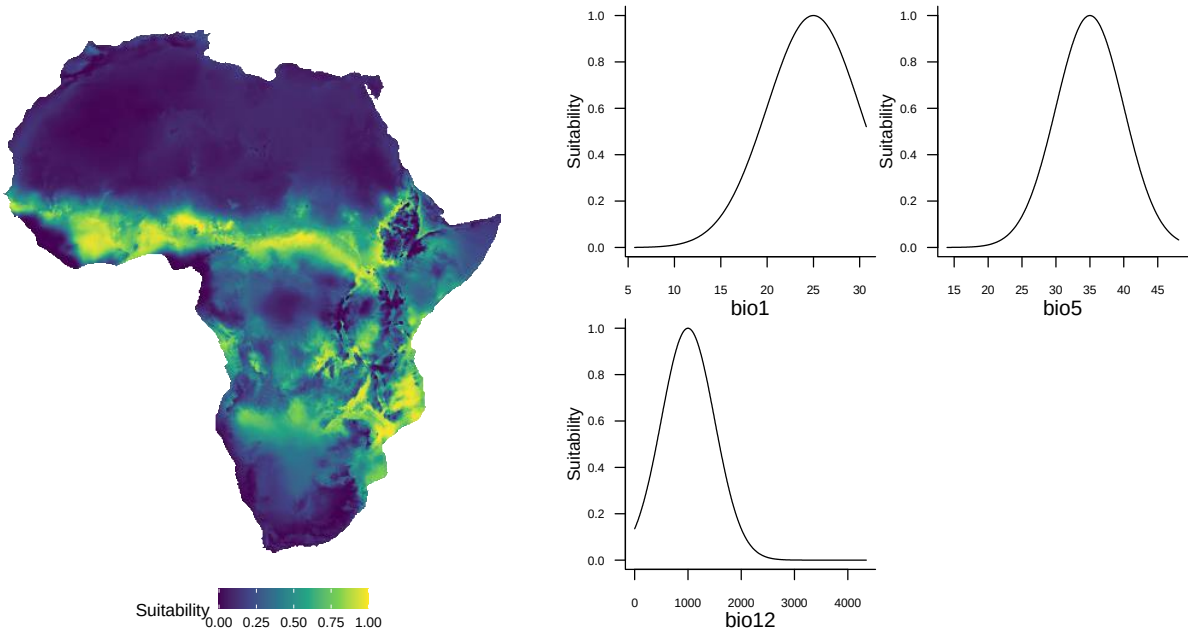

Figure B-2. Environmental suitability and response curves of the simulated virtual species.

B.1.2.2 Explanatory variables selection

Now, with some necessary background knowledge (assuming), we select explanatory variables BIO1, BIO5, BIO12, and three other unrelated random variables (var1-3) for modelling. var1-3 is generated by Gaussian random fields neutral landscape model with function `NLMR::nlm_gaussianfield`.

```
library(NLMR)

# Select variables for modelling
env_vars <- bios %>% slice('band', c(1, 5, 12))

# Make pseudo landscape as other unrelated environmental variables
set.seed(123)
mag_vars <- c(5, 8, 11)
gaussian_field <- function(mag_var) {
  st_as_stars(nlm_gaussianfield(
    nrow = 435, ncol = 414,
    mag_var = mag_var, rescale = FALSE))}
vars <- do.call(c, lapply(mag_vars, gaussian_field))
```

```
# Formatting the vars
names(vars) <- paste0('var', 1:length(mag_vars))
st_dimensions(vars)[1] <- st_dimensions(env_vars)[1]
st_dimensions(vars)[2] <- st_dimensions(env_vars)[2]
vars <- vars * (env_vars %>% slice(band, 1) %>%
                  setNames('mask') %>%
                  mutate(mask = ifelse(is.na(mask), NA, 1)))

# Put them together
env_vars <- c(env_vars, vars %>% merge(name = "band"), along = 3)
rm(vars, mag_vars)
```

B.1.2.3 Pre-analysis of environmental variables and occurrences

As a first step in modelling, it is always helpful to understand the model's inputs, such as the relationships between variables and the quality of the occurrence data. The `ITSDM` provides two functions to enable these assessments: `dim_reduce` and `suspicious_env_outliers` (Table 1 in main text). `dim_reduce` can examine the correlations between variables and reduce the dimension with a user-defined correlation threshold. `suspicious_env_outliers` can detect suspicious environmental outliers in occurrence observations. It is usually unnecessary to use `suspicious_env_outliers` when using iForest, but it can be beneficial to reduce potential sampling errors for other models that are more sensitive to outliers.

B.1.2.3.1 Function dim_reduce

We can set a threshold value of Pearson correlation (`threshold`) and optionally se t preferred variables (`preferred_vars`) for function `dim_reduce`:

```
# Exam and reduce the correlations with a user defined threshold between vari
ables
vars_uncor <- dim_reduce(
  env_vars, threshold = 0.7,
  preferred_vars = "bio1")

vars_uncor
```

```
## Dimension reduction
## Correlation threshold: 0.7
## Original variables: bio1, bio5, bio12, var1, var2, var3
## Variables after dimension reduction: bio1, bio5, bio12, var1, var2, var3
## ============================================================================
## Reduced correlations:
##        bio1  bio5 bio12  var1  var2  var3
## bio1   1.00  0.66 -0.04 -0.03 -0.04  0.05
## bio5   0.66  1.00 -0.59  0.07 -0.01 -0.01
## bio12 -0.04 -0.59  1.00 -0.05  0.01  0.05
## var1  -0.03  0.07 -0.05  1.00  0.11  0.04
## var2  -0.04 -0.01  0.01  0.11  1.00  0.04
## var3   0.05 -0.01  0.05  0.04  0.04  1.00
```

According to the test result, the environmental variables are relatively independent. It is worth noticing that BIO1 and BIO5 have a correlation of 0.66. In the virtual species case, the variables are not strongly correlated as intended, so it is not necessary to remove any variables. For other cases, we can extract `vars_uncor$img_reduced` from the result as a new variable stack with strongly correlated variables removed.

B.1.2.3.2 Function suspicious_env_outliers

`suspicious_env_outliers` function takes a `data.frame` of occurrences and a `stars` of environmental variables and other optional arguments. Because initially, we are not sure about the potential outliers, we recommend using `rm_outliers = FALSE` (which is the default) in function `suspicious_env_outliers` not to remove outliers. After careful checking (e.g. with Figure B-3), we may re-run `suspicious_env_outliers` with `rm_outliers = TRUE` or may selectively delete some of the detected outliers by hand.

```
# Detect suspicious environmental outliers in occurrence dataset
set.seed(123)
occ_outliers <- suspicious_env_outliers(
  po_df, variables = env_vars,
  z_outlier = 5, outliers_print = 0,
  visualize = FALSE)

# Plot
```

```
plot(occ_outliers) +
  theme(plot.margin = margin(rep(0, 4)))
```

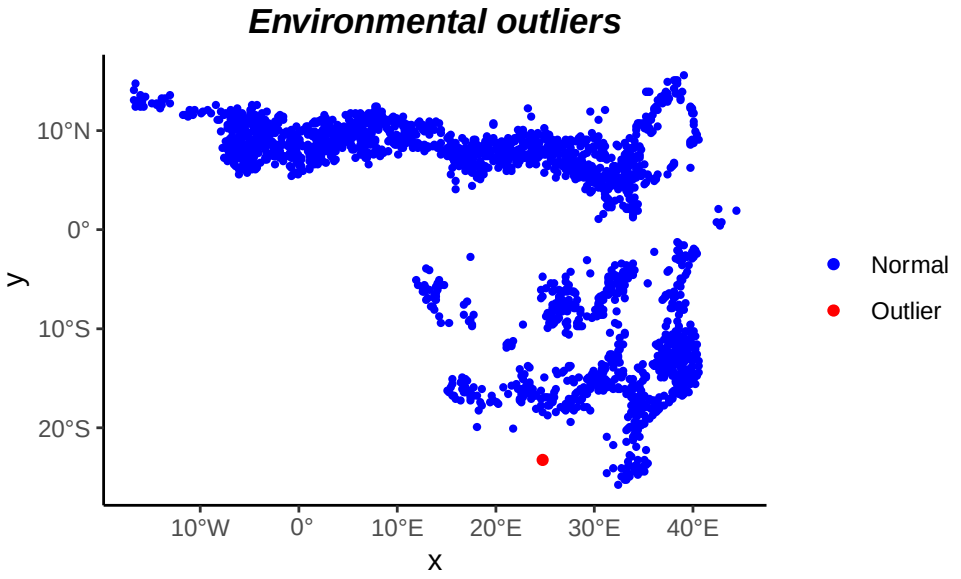

Figure B-3. Outlines detected in the occurrence dataset.

B.1.2.4 Model fitting and insighting with `itsdm` package

B.1.2.4.1 Format dataset for `itsdm`

`format_observation` provided by `itsdm` can help users to format the observations quickly to fit into `itsdm` workflow. A few options need to be set, such as the column name of coordinate x (`x_col`), of coordinate y (`y_col`), and of observation (`obs_col`). We can also set the type of data (`obs_type`) to obtain. For instance, we format the virtual observations to presence-only format and split it into training (70%) and evaluation (30%):

```
# Format the occurrences
obs_formatted <- format_observation(
  obs_df = po_df %>% mutate(obs = 1),
  split_perc = 0.3, x_col = "x", y_col = "y",
  obs_col = "obs", obs_type = "presence_only")
```

```
# Check it
print(obs_formatted)
```

```
## Formatted occurrence observations
##   Training observations:
##     Type: presence_only
##     No. of observations: 1400
##   Evaluation observations:
##     Type: presence_only
##     No. of observations: 600
```

B.1.2.4.2 Fit the iForest SDM

After the data pre-processing steps, a model is fit to the provided set of occurrence points (`obs`) and corresponding environmental variables (`env_vars`) with a few other settings (Table S1-1):

```
# Create an Extended isolation forest
obs <- obs_formatted$obs
eval <- obs_formatted$eval
mod <- isotree_po(
  obs = obs,
  obs_ind_eval = eval,
  variables = env_vars,
  sample_size = 0.8,
  ndim = 2)

# Check the result
# It will print out all relevant information
print(mod)
```

```
## ============================================================
## Species distribution model:
## Extended Isolation Forest model
## Splitting by 2 variables at a time
## Consisting of 100 trees
## Numeric columns: 6
## Size: 5.66 MiB
## Variables are: bio1, bio5, bio12, var1, var2, var3.
## Use independent test? FALSE.
## Has marginal responses? TRUE.
## Has independent responses? TRUE.
## Has Shapley value based responses? TRUE.
## Has spatial responses? TRUE.
```

```
## Has variable analysis? TRUE.
## ============================================================
## Model evaluation:
## [Training dataset]:
## ==================================
## Presence-only evaluation:
## CVI with 0.25 threshold:      0.176
## CVI with 0.5 threshold:       0.693
## CVI with 0.75 threshold:      0.635
## CBI:                          0.998
## AUC (ratio)                   0.924
## ==================================
## Presence-background evaluation:
## Sensitivity:                  0.956
## Specificity:                  0.800
## TSS:                          0.756
## AUC:                          0.932
## Similarity indices:
## Jaccard's similarity index:   0.797
## Sørensen's similarity index:  0.887
## Overprediction rate:          0.173
## Underprediction rate:         0.044
## [Test dataset]:
## ==================================
## Presence-only evaluation:
## CVI with 0.25 threshold:      0.176
## CVI with 0.5 threshold:       0.689
## CVI with 0.75 threshold:      0.618
## CBI:                          0.991
## AUC (ratio)                   0.922
## ==================================
## Presence-background evaluation:
## Sensitivity:                  0.953
## Specificity:                  0.808
## TSS:                          0.762
## AUC:                          0.936
## Similarity indices:
## Jaccard's similarity index:   0.800
## Sørensen's similarity index:  0.889
## Overprediction rate:          0.167
## Underprediction rate:         0.047
## ============================================================
## Variable importance:
## Just show SHAP result, use print or plot of variable_analysis to see all.
## SHAP (mean(|Shapley value|))
## [Training dataset]:
## bio12 : ############################################## 0.039
## bio5  : ############################## 0.024
```

```
## bio1  : ############################ 0.023
## var1  : ############## 0.012
## var3  : ############# 0.011
## var2  : ############ 0.01
## [Test dataset]:
## bio12 : ############################################ 0.039
## bio5  : ############################## 0.025
## bio1  : ############################ 0.023
## var1  : ############## 0.012
## var3  : ############## 0.012
## var2  : ############ 0.01
```

Both presence-only and presence-absence (background) evaluation metrics are implemented (see the description of all metrics in Appendix S2). The print method displays all metrics, summarizing the performance of the model, and plot shows those that can be effectively visualized (Figure B-4), like the continuous predicted-to-expected (P/E) curve (Boyce et al., 2002; Hirzel et al., 2006).

```
plot(mod$eval_test))
```

```
# Diagnostic plots according to test dataset
plot(mod$eval_test)
```

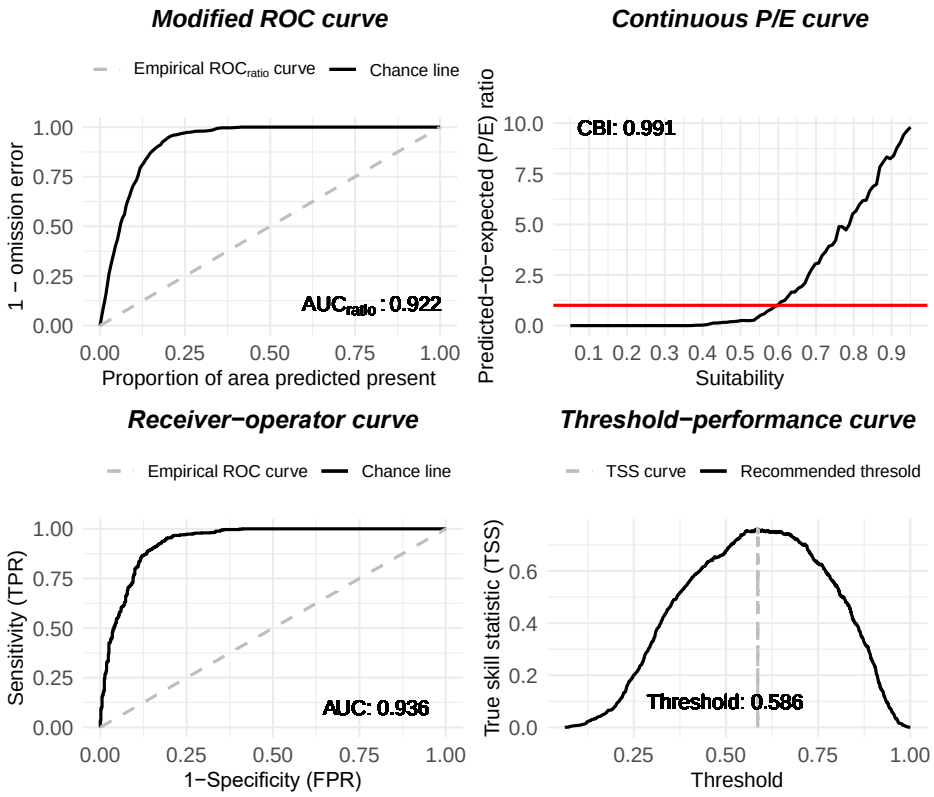

Figure B-4. The model diagnostic shown by plot function (made by `plot(mod$eval_test)`)

A high Continuous Boyce Index (CBI) (Hirzel et al., 2006), Area Under the ROC Curve (AUC) (Hanley & others, 1989), and modified AUC (AUC_{ratio}) (Appendix S2) in Figure B-4 reflected that the model had done adequately. And the model was calibrated well with higher species density in areas with higher relative habitat suitability as reflected by a consistently growing P/E curve (Figure B-4).

B.1.2.5 Model explanation

In SDM, model explanation is also very important to get ecological insights from models. In package ITSDM, model explanation includes analysis of variable importance analysis and variable response (both non-spatially and spatially).

B.1.2.5.1 Variable importance

Variable importance could both help to improve the model quality and diagnose the relationships between species and environmental variables:

```
plot(mod$variable_analysis)
```

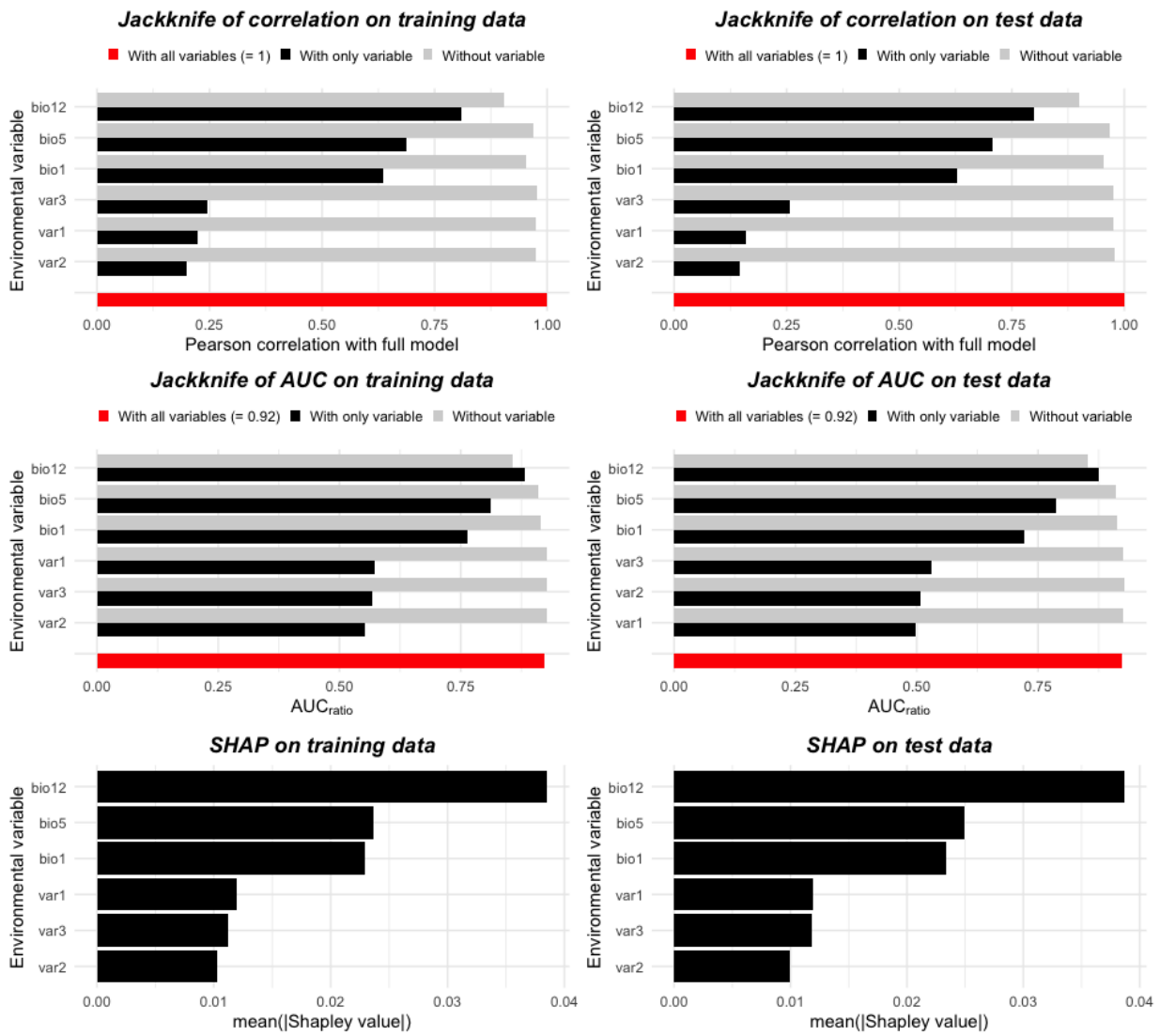

Figure B-5. Variable importance analysis of species distribution model for the virtual species.

As shown in Table 1 in the main text, multiple methods are available for model explanation. Variable importance (calculated by `variable_analysis`) is diagnosed by the

Jackknife test with Pearson correlation between partial prediction and full prediction using all variables, Jackknife test with metrics AUC_{ratio} (see Appendix S2), and Shapley values technique. For instance, the variable analysis for our virtual species case is shown in Figure B-5. According to all three methods, variables bio12, bio5, and bio1 have much higher importance than var 1 through var 3, as intended. In addition, the similarity in the values for these metrics for both the training and test dataset indicates that the model is well-generalized.

B.1.2.5.2 Variable responses curves

Response analysis of environmental variables is critical in SDM. ITSDM employs several methods to generate response curves. Like many other SDMs (e.g., Maxent), the marginal response curve is coded to show how prediction varies as each variable changes when all other environmental variables are at their average sample value (Elith et al., 2005). The independent response curve shows how prediction changes as each environmental variable are varied singly. The Shapley value-based response curve conveys how prediction is pushed away from the average prediction across the whole training dataset (Molnar, 2020). The Shapley values also allow users to diagnose the correlation between two variables of both raw values and contributions to the model results (Figure 2-4 in the main text).

Similar `plot` functions are applied to marginal response plots, independent response plots, and Shapley values-based response plots. Let's see an example with marginal response curves:

```
# Take marginal response as an example
# Generate marginal response curves
# if choosing response = FALSE in isotree_po
marginal_responses <- marginal_response(
  model = mod$model,
  var_occ = mod$vars_train,
  variables = mod$variables)
```

```
# Plot marginal response curves
# Here mod$marginal_responses is equal to marginal_responses
plot(mod$marginal_responses,
     target_var = c('bio1', 'bio12')) +
  theme_classic()
```

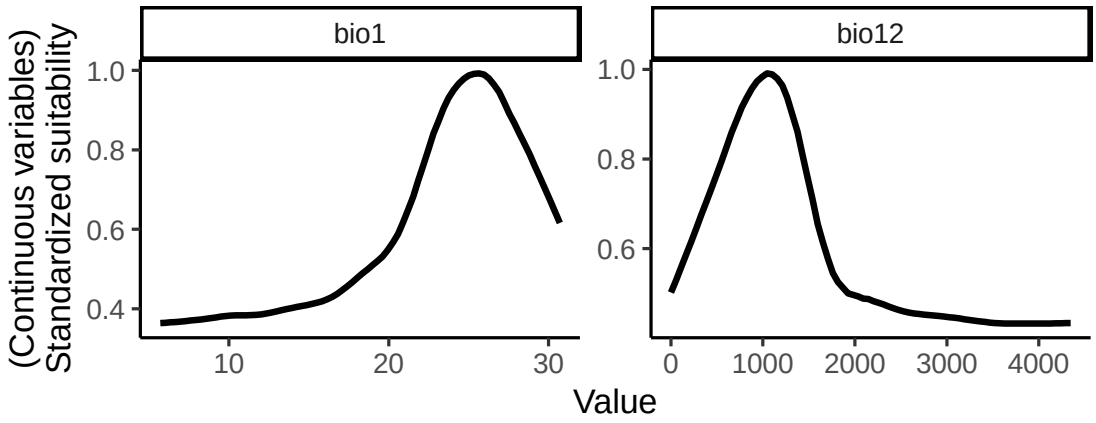

Figure B-6. Marginal response curves of bio1 and bio12 for the virtual species.

B.1.2.5.3 Spatial response maps

As described in the main text, `itsdm` has the function (`spatial_response`) to analyze the spatial response maps. The function performs the same calculation as response curves, but for all pixels in environmental maps. Therefore, it includes marginal response maps, independent response maps, and Shapley values-based response maps (Figure B-7 and Figure 2-5 in the main text):

```
# Make spatial response maps with all three methods
# Make sure to set a non-zero shap_nsim
full_spatial_responses <- spatial_response(
  model = mod$model,
  var_occ = mod$vars_train,
  variables = mod$variables,
  shap_nsim = 10)

# Check bio12
plot(full_spatial_responses, target_var = 'bio12')
```

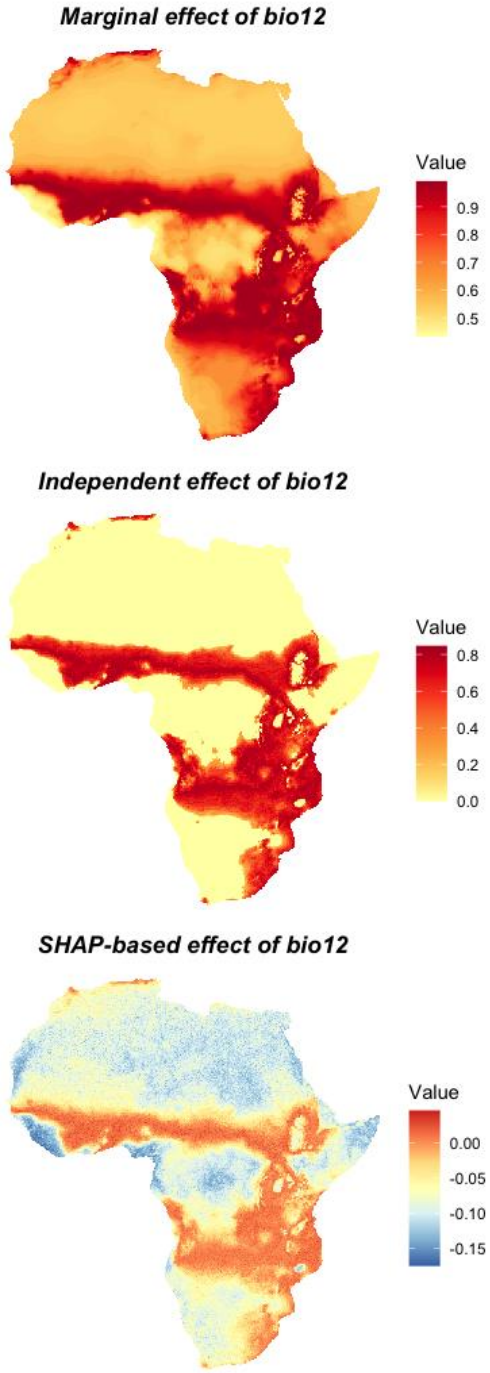

Figure B-7. Spatial response map of variable bio12 for the virtual species.

All three maps consistently indicate that bio12 is more critical in some areas. Shapely value-based effect further proves bio12 contributes minimally over some areas even though it is the most vital environmental covariate diagnosed in variable analysis.

B.1.2.6 Post-analysis

B.1.2.6.1 Variable contribution analysis

After exploring variable importance and response plots, we can re-select variables and recreate a more accurate model. When the final model is done, sometimes we would like to diagnose some specific observations. `variable_contrib` could help to diagnose the variable contribution to specific observations (Figure B-8). For example:

```
# Diagnose variable contribution
var_contrib <- variable_contrib(
  mod$model,
  mod$vars_train,
  mod$vars_test %>% slice(1:4))

# Check the results
plot(var_contrib, plot_each_obs = T)
```

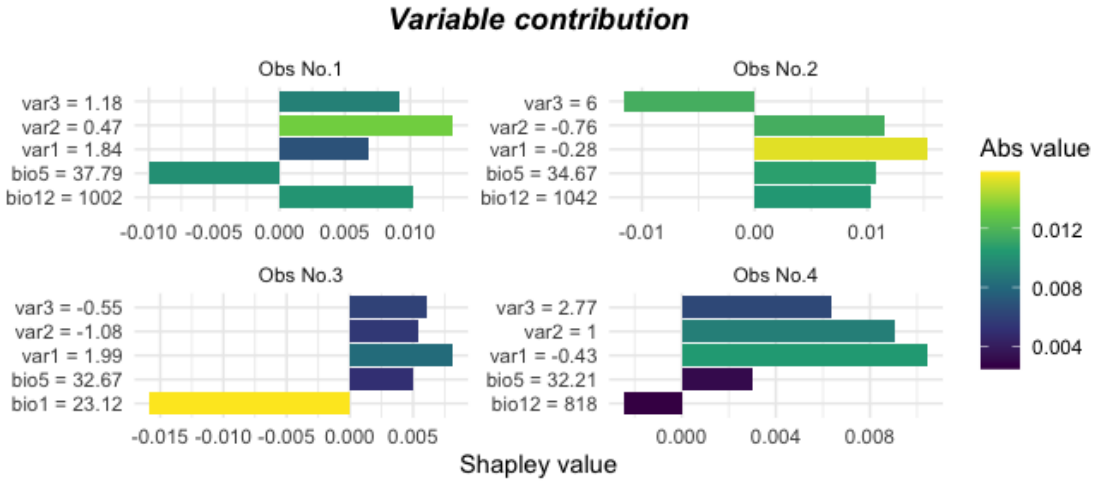

Figure B-8. Variable contributions to the modelled suitability of selected occurrence observations.

B.1.2.6.2 Convert to presence-absence

Additionally, function `convert_to_pa` can be used to convert the environmental suitability map to probability of occurrence and presence-absence map:

```
# Logistic conversion
pa_log <- convert_to_pa(
  mod$prediction, method = 'logistic',
  beta = 0.5, alpha = -.05)
```

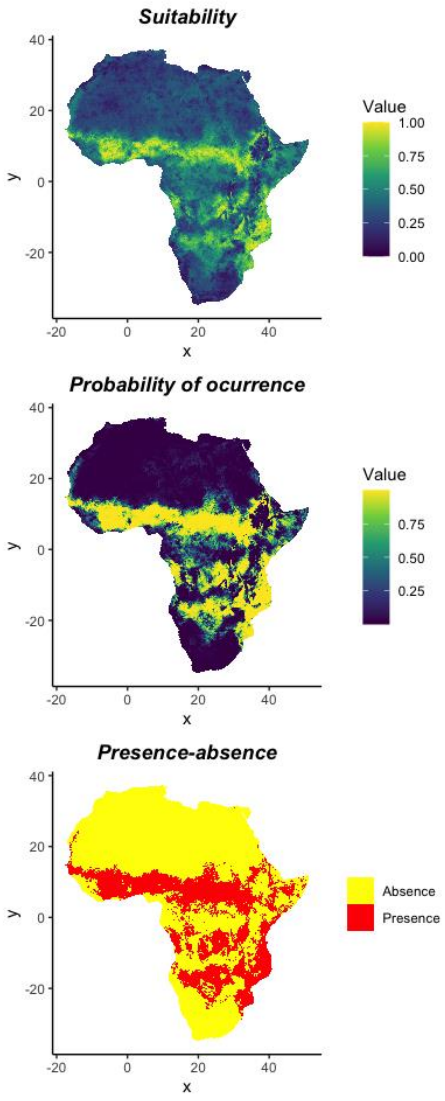

Figure B-9. Presence-absence conversion maps.

B.1.3 Scripts for other examples in the manuscript

All other scripts and data used but not shown in this document are available via Open Science Framework (OSF): https://osf.io/8mc4e/:

- **data**: This folder contains simulated virtual species for model comparison and the evaluation metrics of all models for comparison.

- **scripts**: This folder contains scripts to generate virtual species, run models, and make figures for model comparison. It also hosts the script for changing environment analysis introduced in section 2.5.2.

References

Boyce, M. S., Vernier, P. R., Nielsen, S. E., & Schmiegelow, F. K. A. (2002). Evaluating resource selection functions. *Ecological Modelling*, *157*(2–3), 281–300. https://doi.org/10.1016/S0304-3800(02)00200-4

Cortes, D. (2022). *isotree: Isolation-Based Outlier Detection*. https://CRAN.R-project.org/package=isotree

Elith, J., Ferrier, S., Huettmann, F., & Leathwick, J. (2005). The evaluation strip: A new and robust method for plotting predicted responses from species distribution models. *Ecological Modelling*, *186*(3), 280–289. https://doi.org/10.1016/j.ecolmodel.2004.12.007

Hanley, J. A. & others. (1989). Receiver operating characteristic (ROC) methodology: The state of the art. *Crit Rev Diagn Imaging*, *29*(3), 307–335.

Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C., & Guisan, A. (2006). Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, *199*(2), 142–152. https://doi.org/10.1016/j.ecolmodel.2006.05.017

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com. https://christophm.github.io/interpretable-ml-book/

B.2 Evaluation metrics

B.2.1 Presence-only evaluation

When absence data are unavailable, half of the confusion matrix is missing in presence-only data, which rules out the standard statistics such as receiver operating characteristic curve (ROC) and Kappa. Several indices have then been proposed to evaluate the presence-only SDMs. In this paper, we used Contrast Validation Index (CVI), Continuous Boyce Index (CBI), and AUC_{ratio} based on the modified Receiver Operating Characteristic (ROC) curve (Hanley & others, 1989; Peterson et al., 2008).

CVI is a threshold-dependent evaluator, which is the proportion (A_p) of presences falling above a fixed HS threshold (e.g. 0.5) minus the proportion (A_g) of cells in the whole area above this fixed HS threshold (Hirzel et al., 2006; Hirzel & Arlettaz, 2003; Santika et al., 2019). This paper used 0.25, 0.5, and 0.75 as the threshold, respectively. The values of CVI vary from 0 to $(1 - A_g)$.

The predicted-to-expected ratio (P/E) curve and the continuous Boyce index (CBI) are designed to evaluate the presence-only model in a threshold-independent way. Bins of habitat suitability were chosen by a moving window of fixed size. For each bin i, P/E ratio was calculated by dividing predicted frequency (P_i, Equation 1) and expected frequency (E_i, Equation 2) (Boyce et al., 2002; Giralt Paradell et al., 2019; Hirzel et al., 2006).

$$P_i = \frac{p_i}{\sum_{j=1}^{b} p_j} \tag{B-1}$$

$$E_i = \frac{a_i}{\sum_{j=1}^{b} a_j} \tag{B-2}$$

where p_i is the number of presences predicted by the model to fall in the habitat suitability bin i and $\sum_{j=1}^{b} p_j$ is the total number of presences; a_i is the number of grid cells belonging to habitat suitability bin i and $\sum_{j=1}^{b} a_j$ is the overall number of grid cells.

The CBI is the Spearman correlation between P/E ratios and bin index i. If the index is lower than 0, the model is not reliable. In contrast, if the index is higher than 0, it means the model is reliable. The closer to 1, the more reliable the model is (Giralt Paradell et al., 2019; Hirzel et al., 2006; Santika et al., 2019).

A ROC curve is generated by plotting the proportion of correctly predicted presence on the y-axis against one minus the proportion of correctly predicted absence on the x-axis for all thresholds. The area under the ROC curve (AUC) is a threshold-independent evaluator of model performance, needing both presence and absence data. Multiple approaches have tried to adapt ROC to presence-only data. Peterson et al. (2008) modified AUC by plotting the proportion of correctly predicted presence against the proportion of presences falling above a range of thresholds against the proportion of cells of the whole area falling above the range of thresholds. This is called AUC_{ratio} in itsdm.

B.2.2 Presence-background evaluation

Even though the presence-background evaluation metrics in SDM are repeatedly criticized for many different reasons. In itsdm, we still employed several presence-absence evaluation metrics by extracting a few background points as pseudo absence. Because these

metrics are widely used despite known issues, partially because these metrics have been predominance due to good discriminating capacity. In addition, some SDMs rely on pseudo absence/background points for modeling and evaluation. Using presence-background metrics could provide comparable model evaluation for tasks such as model ensemble. These metrics include sensitivity, specificity, true skill statistic (TSS), threshold-performance curve driven by TSS, ROC, and the AUC. Leroy et al. (2018) argued that similarity/F-measures are robust evaluation metrics to discriminate SDMs without quality presence-absence data. According to their recommendation, we thus include Jaccard's similarity index, Sørensen's similarity index (F-measure), Overprediction rate, and Underprediction rate (Leroy et al., 2018) in itsdm. Assume we get true positive (TP), true negative (TN), false positive (FP), and false negative (FN), these discrimination metrics are calculated as follows (Fielding & Bell, 1997; Jaccard, 1908; Leroy et al., 2018; Li & Guo, 2013; Márcia Barbosa et al., 2013; Sorensen, 1948):

$$sensitivity\ (TPR) = TP/(TP + FN) \tag{B-3}$$

$$specificity\ (TNR) = TN/(TN + FP) \tag{B-4}$$

$$TSS = sensitivity + specificity - 1 \tag{B-5}$$

$$Jaccard's\ similarity\ index\ =\ TP/(FN\ +\ TP\ +\ FP) \tag{B-6}$$

$$Sørensen's\ similarity\ index\ =\ 2TP/(FN\ +\ 2TP\ +\ FP) \tag{B-7}$$

$$overprediction\ rate\ =\ FP/(TP\ +\ FP) \tag{B-8}$$

$$underprediction\ rate\ =\ FN/(TP\ +\ FN) \tag{B-9}$$

The threshold-performance curve is generated by plotting TSS on the y-axis against the threshold values on the x-axis. An optimal cutoff threshold value is generated and used to calculate all other metrics.

References

Boyce, M. S., Vernier, P. R., Nielsen, S. E., & Schmiegelow, F. K. A. (2002). Evaluating resource selection functions. *Ecological Modelling*, *157*(2–3), 281–300. https://doi.org/10.1016/S0304-3800(02)00200-4

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, *24*(1), 38–49.

Giralt Paradell, O., Díaz López, B., & Methion, S. (2019). Modelling common dolphin (Delphinus delphis) coastal distribution and habitat use: Insights for conservation. *Ocean & Coastal Management*, *179*, 104836. https://doi.org/10.1016/j.ocecoaman.2019.104836

Hanley, J. A. & others. (1989). Receiver operating characteristic (ROC) methodology: The state of the art. *Crit Rev Diagn Imaging*, *29*(3), 307–335.

Hirzel, A. H., & Arlettaz, R. (2003). Modeling habitat suitability for complex species distributions by environmental-distance geometric mean. *Environmental Management*, *32*(5), 614–623.

Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C., & Guisan, A. (2006). Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, *199*(2), 142–152. https://doi.org/10.1016/j.ecolmodel.2006.05.017

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, *44*, 223–270.

Leroy, B., Delsol, R., Hugueny, B., Meynard, C. N., Barhoumi, C., Barbet-Massin, M., & Bellard, C. (2018). Without quality presence–absence data, discrimination metrics such as TSS can be misleading measures of model performance. *Journal of Biogeography*, *45*(9), 1994–2002.

Li, W., & Guo, Q. (2013). How to assess the prediction accuracy of species presence-absence models without absence data? *Ecography*, *36*(7), 788–799. https://doi.org/10.1111/j.1600-0587.2013.07585.x

Márcia Barbosa, A., Real, R., Muñoz, A.-R., & Brown, J. A. (2013). New measures for assessing model equilibrium and prediction mismatch in species distribution models. *Diversity and Distributions*, *19*(10), 1333–1338. https://doi.org/10.1111/ddi.12100

Peterson, A. T., Papeş, M., & Soberón, J. (2008). Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling*, *213*(1), 63–72. https://doi.org/10.1016/j.ecolmodel.2007.11.008

Santika, T., Hutchinson, M. F., & Wilson, K. A. (2019). *The effects of data adequacy and calibration size on the accuracy of presence-only species distribution models* [Preprint]. Ecology. https://doi.org/10.1101/775700

Sorensen, T. A. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biol. Skar.*, *5*, 1–34.

Appendix C
Appendix to Chapter 3

C.1 Extended materials and methods

C.1.1 High-resolution land cover mapping in Tanzania

Using a similar method proposed by Song et al. (2023), Sentinel-1 time series (Torres et al., 2012), and PlanetScope NICFI basemaps (*Norway's International Climate and Forest Initiative (NICFI)*, 2020), we created a land cover map over Tanzania with 4.77-meter spatial resolution, which includes 8 main land cover types: Cropland, dense tree, Shrubland, Grassland, Water, Bareland, Built-up, and Wetland. This map was made by running three processes: multi-source land cover products ensemble, pixel-wise Random Forest modeling, and image-wise U-Net modeling. To create synthetic labels, several existing land cover products (Buchhorn et al., 2020; Congalton et al., 2017; Xu et al., 2019) were combined into a consensus land cover map, which can provide a reliable initial set of labels but only provide partial coverage of the study area. To fill the missing data in the resulting synthetic labels, we trained a Random Forest model to fill the gaps for a number of randomly selected tiles, with predictors of temporal features extracted from Sentinle-1 time series, 4 bands (R,G, B and NIR) of PlanetScope imagery, and additional derived vegetation indices. We then visually assessed the gap-filled labels and edited them when necessary. A U-Net model was eventually trained by these labels generated by RF and edited by humans. Unlike the method in Song et al. (2023), we used a regular balanced cross entropy loss function to train the U-Net to simplify the workflow. Additionally, considering the

vast size of the study area, we split the whole country into 5 zones based on agroecological zones (Philipo et al., 2021). A skeleton model was trained using training labels of the entire country first and then fine-tuned to each zone.

We combined areas classified as grassland, shrubland, and wetland to represent the most commonly used habitat of African savanna elephants. The original map has a spatial resolution of 4.77 m. Zonal or focal statistics on 4.77 m were used to aggregate land cover conditions to any scales used in this study, so there can be landscape information even for a scale of 1 km.

C.1.2 Multi-scale species distribution modeling

The species distribution modeling (SDM) is a complex process that involves scale analysis, hyperparameter tuning, and variable selection. Due to the length limitation of the main text, here we provide an extended description of the modeling process, including data processing, model creation, and evaluation.

C1.2.1 Landscape metrics

In accordance with section 3.3.3 in the main text, we employed a land cover map with a resolution of 4.77 m to create a list of landscape metrics to train both regional and landscape scale SDMs. We calculated all landscape metrics using the R package *landscapemetrics*. Here, we present all metrics utilized in the pre-analysis and final models.

C1.2.1.1 Class-level

The contiguity index (CONTIG) is a Shape metric, which measures the spatial connectedness (or contiguity) of cells in patches (LaGro, 1991; McGarigal et al., 2012). Class-

level CONTIG_MN summarizes the mean of each patch belonging to the target class. It is calculated as follows:

$$CONTIG_{MN} = mean(\frac{\left[\frac{\sum_{r=1}^{z} c_{ijr}}{a_{ij}}\right] - 1}{v - 1}) \quad \text{(C-1)}$$

where c_{ijr} is the contiguity value for pixel r in patch ij. v is the size of the filter matrix. 13 is used for v.

Patch density (PD) is an Aggregation metric, which describes the fragmentation of a class (McGarigal et al., 2012). It calculates the number of patches per 100 hectares:

$$PD = \frac{n_i}{A} \times 10000 \times 100 \quad \text{(C-2)}$$

where n_i is the number of patches of class i, and A is the total landscape area in square meters.

Edge density (ED) is an Area and Edge metric, which indicates the configuration of the landscape (McGarigal et al., 2012). It is the sum of all edges of class i in meters per hectare in relation to the entire landscape area:

$$ED = \frac{\sum_{k=1}^{m} e_{ik}}{A} \times 10000 \quad \text{(C-3)}$$

where e_{ik} is the total edge length in meters and A is the total landscape area in square meters.

The mean of patch area (AREA_MN) is an area and edge metric, which summarizes class i as the mean of all patch areas belong to this class:

$$AREA_{MN} = mean(AREA_{ij}) \quad \text{(C-4)}$$

where $AREA_{ij}$ is the area of each patch in hectares.

Class ratio (or coverage) is the proportion of class i within a selected landscape:

$$Coverage = \frac{a_i}{A} \times 100 \quad \text{(C-5)}$$

where a_i is the area of class i and A is the total landscape area in square meters.

C1.2.1.2 Landscape-level

Patch density (PD) at landscape level describes the fragmentation of the entire landscape area (McGarigal et al., 2012):

$$PD = \frac{N}{A} \times 10000 \times 100 \quad \text{(C-6)}$$

where N is the number of patches, and A is the total landscape area in square meters.

Patch richness density (PRD) is a diversity metric, which measures the number of classes in a landscape. It is one of the simplest diversity and composition metrics, and is calculated as follows:

$$PRD = \frac{m}{A} \times 10000 \times 100 \quad \text{(C-7)}$$

where m is the number of classes and A is the total landscape area in square meters.

Shannon's diversity index (SHDI) is also a diversity metric, which describes the richness and evenness of a landscape (McGarigal et al., 2012; Shannon, 1948). It is widely used in biodiversity and ecology because it takes both the number and abundance of classes into account:

$$SHDI = -\sum_{i=1}^{m} (P_i \times lnP_i) \quad \text{(C-8)}$$

where P_i is the proportion of class i.

C1.2.2 Evaluation metrics

In this study, we used R package *itsdm* for species distribution modeling, which includes model evaluation and result assessment. Detailed calculation of evaluation metrics is recorded in the package documentation and introduction paper. Here we present the metrics used in the analysis of this study.

A receiver operating characteristic curve (ROC) curve is generated by plotting the proportion of correctly predicted presence on the y-axis against one minus the proportion of correctly predicted absence on the x-axis for all thresholds (Fielding & Bell, 1997; Hanley & others, 1989). The area under the ROC curve (AUC) is a threshold-independent evaluator of model performance, needing both presence and absence data. When absence data are unavailable, half of the confusion matrix is missing in presence-only data, which rules out the standard statistics such as ROC and AUC. Multiple approaches have tried to adapt ROC to presence-only data. Peterson et al. (2008) modified ROC (ROC_{ratio}) by plotting the proportion of correctly predicted presence against the proportion of presences falling above a range of thresholds against the proportion of cells of the whole area falling above the range of thresholds. The area under the ROC_{ratio} curve is named AUC_{ratio}.

Other metrics: Sensitivity and F-measure were used in this study by randomly collecting background samples. These metrics are calculated as follows (Fielding & Bell, 1997; Leroy et al., 2018; Márcia Barbosa et al., 2013; Sorensen, 1948) with true positive (TP), true negative (TN), false positive (FP), and false negative (FN):

$$Sensitivity\ (TPR) = TP/(TP + FN) \quad \text{(C-9)}$$

$$F - measure\ =\ 2TP/(FN\ +\ 2TP\ +\ FP) \quad \text{(C-10)}$$

C1.2.3 Regional species distribution modeling

Different environmental variables (Table 3-1 in the main text and Table C-1) used in regional SDM are calculated differently for zonal calculation. Specifically, within the defined extent of each coarse spatial grain, NDVI values were calculated using zonal mean, surface roughness was calculated as the standard deviation of all elevation values, vector-based features (such as roads, rivers, and settlements) were calculated based on density, and landscape metrics were calculated using corresponding functions (see equations in section C.1.2.1) (Ntukey et al., 2022).

We used a 10-fold cross-validation and a combined assessment based on area under the modified Receiver Operating Characteristic curve (AUC_{ratio}), area under the receiver operating characteristic curve (AUC), and Sørensen's similarity index (F-measure) (Hanley & others, 1989; Leroy et al., 2018; Peterson et al., 2008) to evaluate model performance and tune model hyperparameters. The tuned hyperparameters of Isolation Forest include the number of pseudo samples and sub-sampling size, max depth of the tree, the number of features to combine to produce a branch split, and scoring metric of the Isolation Forest model (Cortes, 2022).
We subsequently created models at different scales (10, 5, and 2.5 arc-minutes) with optimal hyperparameters, and evaluated the models with both 10-fold cross-validation and the real presence-only occurrences (Section 3.3.2 in the main text). Because the real occurrences are presence-only and biased and fragmented spatially, we only used Sensitivity ($TP/(TP+FN)$) (Hirzel et al., 2006) as the assessment metric. Sensitivity was calculated with the mean value of the optimal thresholds obtained from 10-fold cross-validation. We selected the optimal scale at

which the model has the best performance as the regional scale and used the model created at that scale as the regional-scale model.

C1.2.4 Landscape species distribution modeling

As described in section 3.3.3.2 in the main text, for landscape scale modeling, we used the Kolmogorov–Smirnov (K-S) distance (Massey Jr, 1951) between the values of variables at occurrence coordinates and the values of all variables across the expert range area to exclude the variables that cannot represent the conditions across suitable habitats in Tanzania. Figure C-1 shows the distribution of these values and their Kolmogorov–Smirnov (K-S) distance. Variables with K-S distance smaller than 0.4 were selected in this step. 0.4 is an arbitrary value that can remove variables that have notable different distributions.

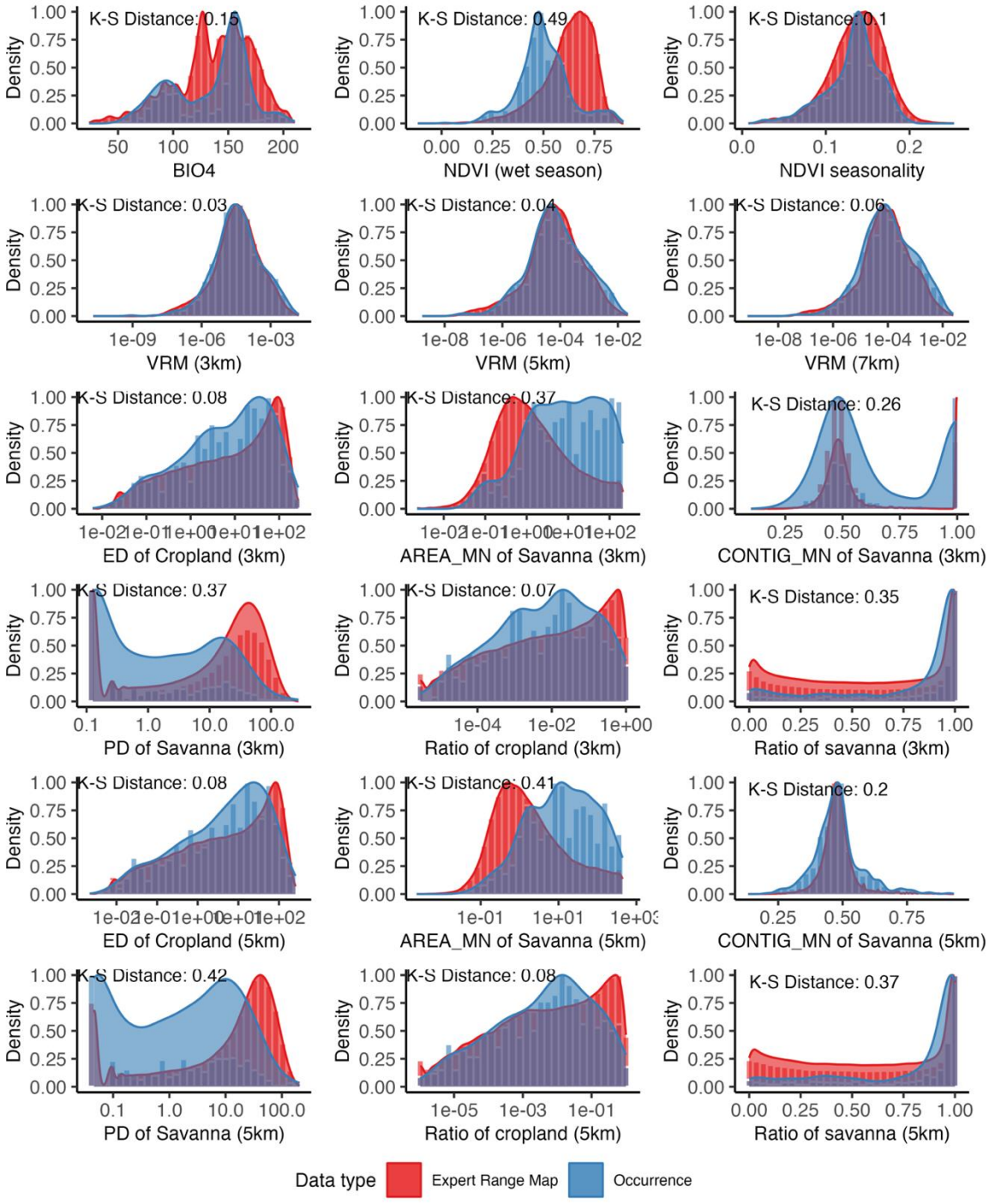

Figure C-1. Kolmogorov–Smirnov (K-S) distances between values of environmental variables extracted by occurrences and expert range map. VRM is Vector Ruggedness Measurement; ED is edge density; AREA_MN is mean of patch area; CONTIG_MN is mean of contiguity index; PD is patch density.

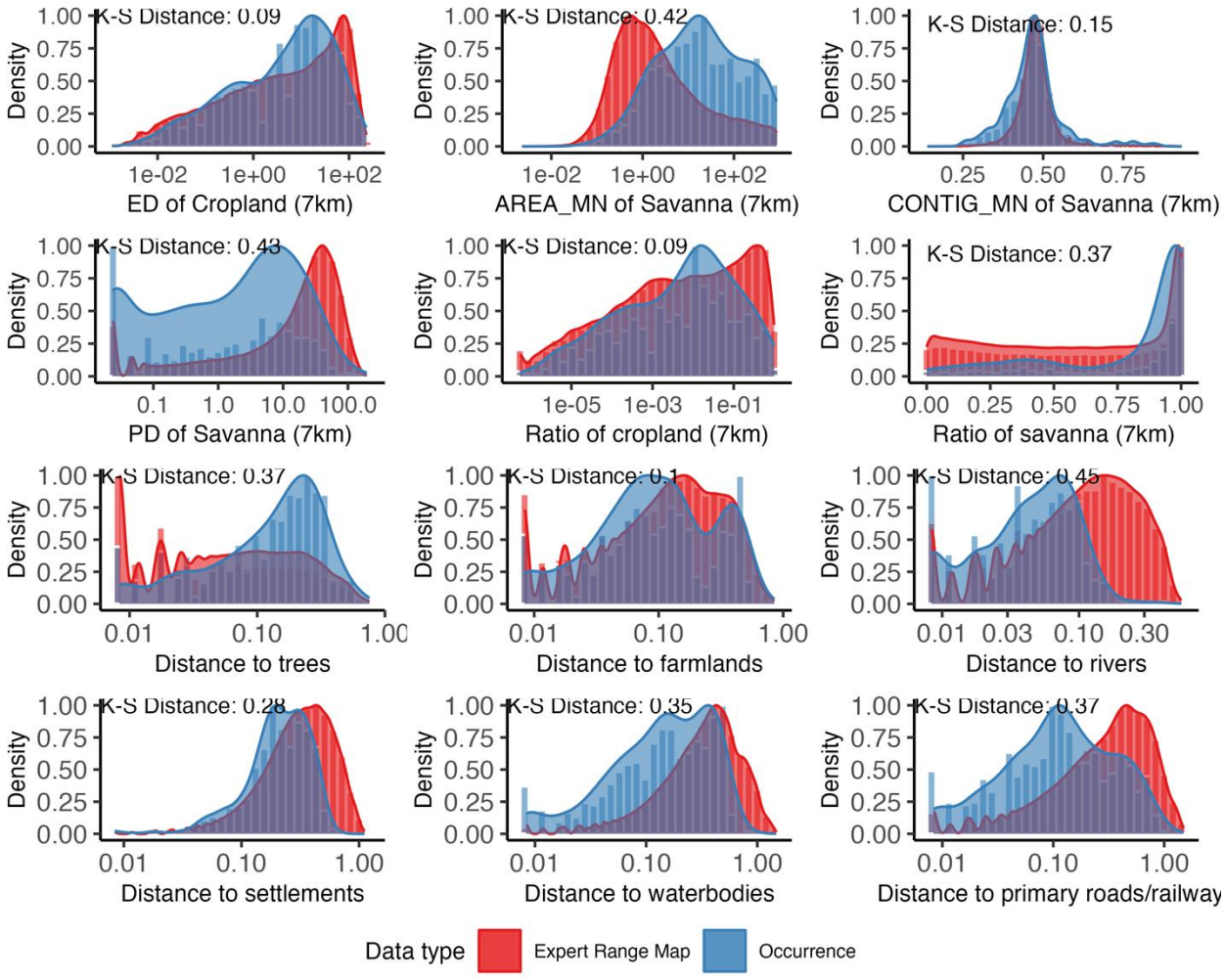

Figure C-1 (Continued). Kolmogorov–Smirnov (K-S) distances between values of environmental variables extracted by occurrences and expert range map. VRM is Vector Ruggedness Measurement; ED is edge density; AREA_MN is mean of patch area; CONTIG_MN is mean of contiguity index; PD is patch density.

For univariate scaling analysis, we created Isolation Forest models with cross-validation to test the model and used the combination of average AUC_{ratio} and AUC as the criterion to assess the model's predictive ability. For each variable, the window size at which the average of AUC_{ratio} and AUC was highest and greater than 0.5 was selected. If the variable with the detected optimal window size was removed from the previous K-S test, the variable with the second optimal window size was used, and so on.

C.1.3 Landscape connectivity

In Scenario A (current condition), we chose pixels (1 km^2) with over 80% cropland coverage and containing more than 400 settlements as areas with intensive human disturbances. Leaving space for settlements, roads, and trees, 80% cropland coverage is a conservative value that can represent a pure cropland landscape (see Figure C-2). After overlapping cities and big towns with settlement density layer, 400 settlements per square kilometer can represent busy human centers, such as towns and cities (Figure C-2). In Scenario B (potential condition), we chose pixels (1 km^2) with over 20% cropland coverage and containing more than 200 settlements as areas with intensive human disturbances, assuming areas with existing human activity will continue to expand. Considering African population will double by 2050, we simply double the settlement density to match the population increase. Of course, this is not the realistic case, but it can give us a simple estimation of population expansion. Similarly, considering the rapid increase of food demand and crop yield in Africa will not increase significantly in a short-term, arable land area will increase to meet the food demand.

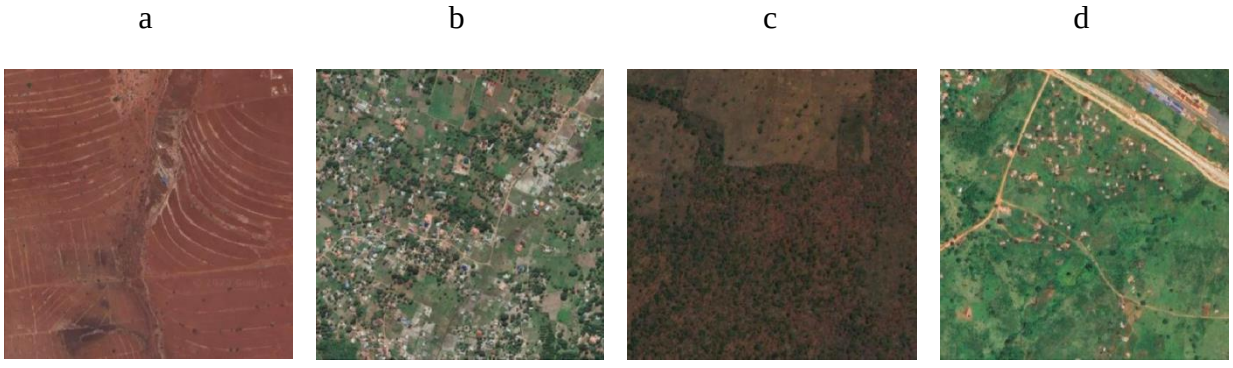

Figure C-2. Examples of pixels (1 km^2) with over 80% cropland coverage (a, 83%), containing more than 400 settlements (b, 465/km^2), with over 20% cropland coverage (c, 22%), and containing more than 200 settlements (d, 237/km^2). Panel a and b represent the landscape that

already have intensive human disturbances, and panel c and d represent the landscape that potentially will become intensively disturbed by human activities.

The primary habitat clusters used for landscape connectivity analysis and the distribution of all involved Protected areas (PAs) in Tanzania is shown in Figure C-3.

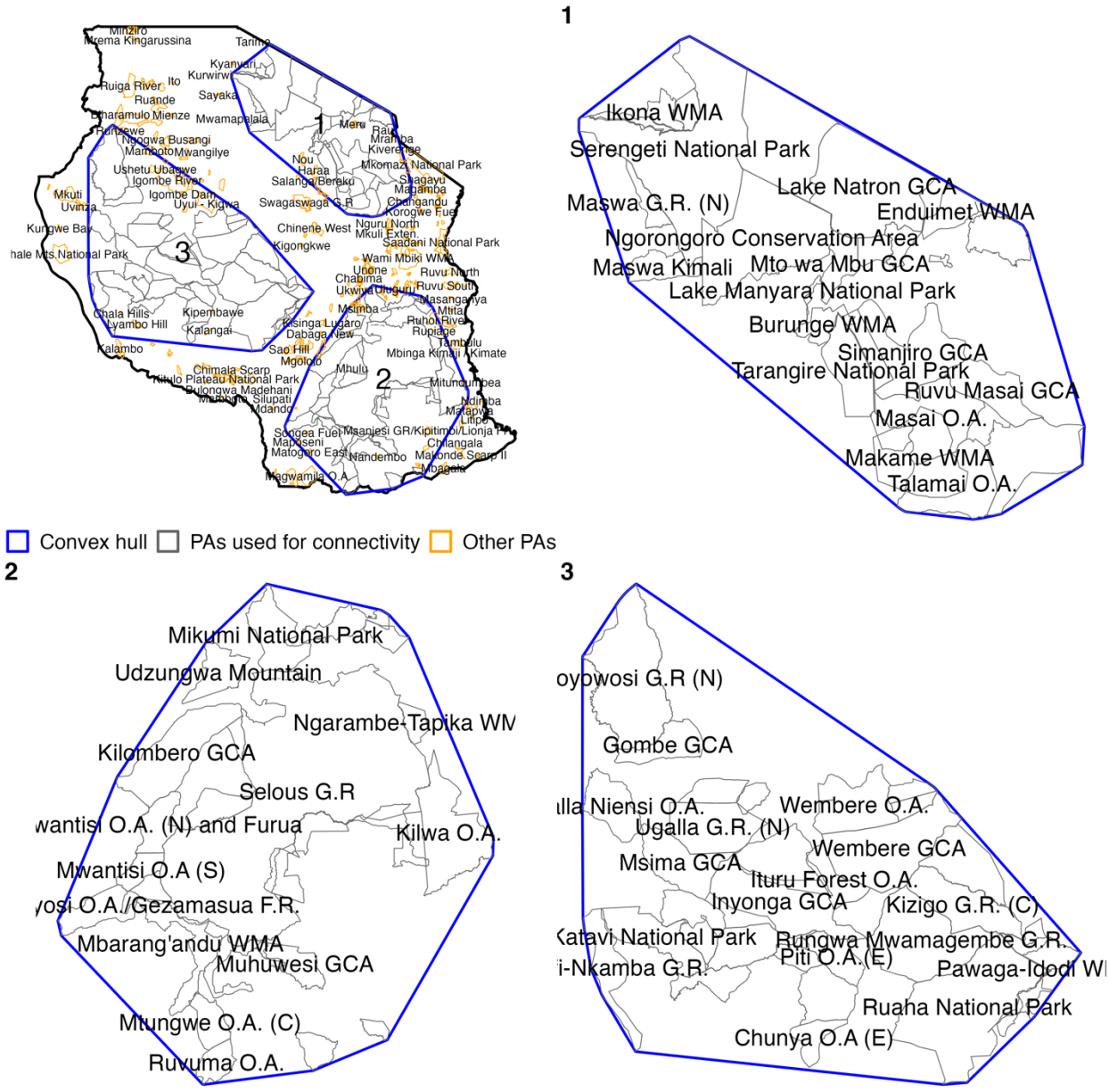

Figure C-3. Three primary habitat clusters used for landscape connectivity analysis and the involved Protected areas (PAs) in Tanzania.

C.2 Extended results and discussions

C.2.1 Species distribution modeling with polygon-based observations

According to previous research (Fourcade, 2016), the number of pseudo occurrences sampled from the expert range map (see section 3.3.2 and 3.3.3.1 in the main text) influences model performance together with modeling scale by controlling the marginal environmental regions. We did the sampling ratio analysis for regional-scale modeling, and the results are shown in Figure C-4. The best sampling ratio to generate pseudo-occurrences is 0.3 at the spatial resolution of 2.5 arc-minutes, 0.2 at 5 arc-minutes, and 0.4 at 10 arc-minutes (Figure C-4). The number of pseudo samples has a consistent effect on AUC_{ratio}, which is a presence-only evaluation metric. With the increase of sampling ratio, AUC_{ratio} rises to a plateau and then fluctuates around the plateau or slightly drops. Presence-background evaluation metrics, such as AUC and F-measure, however, usually respond to the number of pseudo samples along a curve with turning points (Figure C-4). How the curves change depends on the corresponding spatial scale. For instance, in Figure C-4, when the spatial resolution is 2.5 arc-minutes, AUC and F-measure decrease, then increase, and then slightly decrease with the rising of sampling ratio. When the spatial scale is coarse (5 or 10 arc-minutes), AUC and F-measure increase and then decrease. It may suggest that even though presence-background evaluation metrics have notable drawbacks in presence-only species distribution modeling, they are irrefutably useful to distinguish models' predictive power, particularly in hyperparameter tuning. Ideally, both presence-only and presence-background evaluation metrics should be considered to evaluate presence-only SDMs.

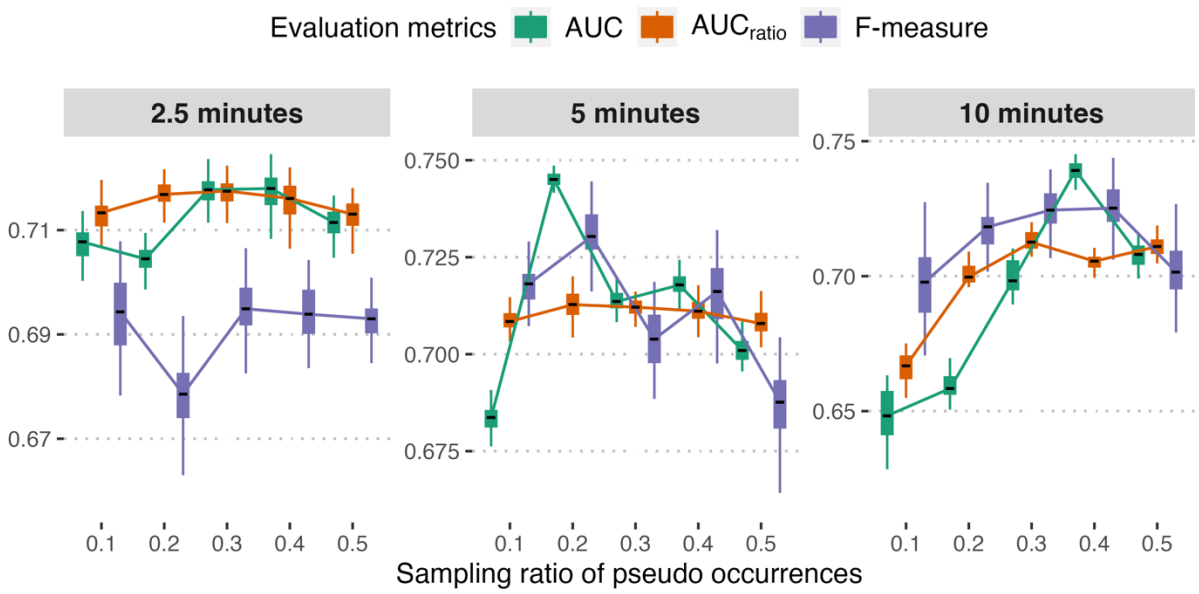

Figure C-4. Effect of sampling ratio of pseudo-occurrences on model performance evaluated by different evaluation metrics (mean and 95% confidence intervals) based on cross validation at different coarse scales (2.5, 5, and 10 arc-minutes).

Unexpectedly but reasonably, the best sampling ratio drops from 0.3 to 0.2 and rises to 0.4 with the increase of spatial resolution from 2.5 to 10 arc-minutes. It may indicate that there is an optimal tradeoff between the number of pseudo-occurrences and spatial resolution on model performance. The curves in Figure C-4 denote that the spatial grain of elephants respond to environmental factors generally at the coarse scale is near or larger than 10 arc-minutes because if the response scale is smaller than 10 arc-minutes, the model performance at the scale of 10 arc-minutes will decrease with an increasing sampling ratio due to the significant marginal impacts of environmental niche. Consequently, 2.5 arc-minutes is too fine to model the environmental suitability of African savanna elephant at coarse scale (Figure C-4) by introducing considerable fine-scale environmental variations and uncertainties. It is also evident in Figure C-5, which shows the performance of models created with optimal sampling ratio of pseudo-occurrences and best hyperparameters at each scale. The low sensitivity calculated on real occurrence dataset at

scale of 2.5 arc-minutes (Figure C-5) specifies poor model generalization. Therefore, 5 and 10 arc-minutes are both suitable scales to model the distribution of African savanna elephants at the coarse scale. The models created at these two scales have a good explanation of the environmental niche at the same scale (Figure C-5A) and a good correspondence to the lower scale (Figure C-5B). Considering slightly better model performance and significantly more detailed representation of the landscape, we selected 5 arc-minutes as the scale for the regional scale modeling (section 3.3.3.1 in the main text).

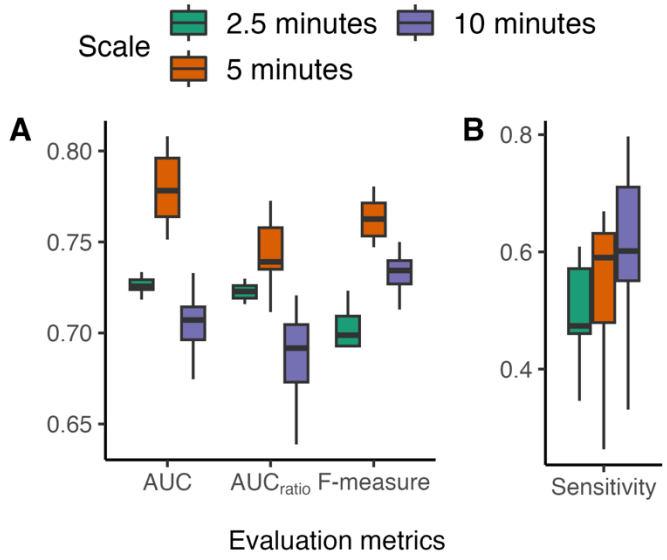

Figure C-5. Comparison of model performance evaluated by different evaluation metrics (mean and 95% confidence intervals) based on cross validation at different coarse scales (2.5, 5, and 10 arc-minutes). A is the evaluation based on cross validation using pseudo-occurrences with the best sampling ratio (section 3.3.3.1 and Figure 3-2). B is the sensitivity calculated by real occurrences with the optimal threshold obtained from cross validation.

C.2.2 Spatial scaling of environmental variables

It is noteworthy that the contribution of many environmental drivers at coarse scale are scale dependent (Figure 3-3). Most of the density-based constraint features including Settlement

density, Road density, and Patch richness density, are more influential at coarser scales. It because these constraints take effect by providing context above the ecological level of elephant distribution in the landscapes (Turner & Gardner, 2015). In contrast, resource features such as NDVI, Tree edge density, and Contiguity index distribution of Savanna, are more important at finer scales. These factors take effect at the ecological level of elephant distribution or below by providing the details needed to explain the distribution (Turner & Gardner, 2015). The contribution of climatic variables commonly does not have strong scale dependence (Figure 3-3). The scaling analysis of variable importance, therefore, can be potentially used to target the most effective grain for each environmental feature at coarse scales to determine species distribution.

Similarly, African elephants' statistical responses to environmental variables are scale dependent, which can be summarized into three types (Figure C-6, 7, & 8). African elephants respond to all climatic variables (BIO1, 4, 7, 12, and 15), NDVI variables (seasonality and seasonal mean), Cropland edge density, Patch density, Ratio of waterbodies, Savanna patch density, and Tree edge density without significant difference among the majority of values across scales (Figure C-6). The seemingly differences in unsuitable values of some variables (e.g. NDVI, NDVI seasonality, and BIO1) mainly result from the different grain size to do zonal calculation (section 3.3.3.1). African elephants respond to Mean of Contiguity index ($CONTIG_{MN}$) of savanna, River density, Road density, and Settlement density with response curves that have same central values but different ranges across scales (Figure C-7). African elephants respond to $CONTIG_{MN}$ of savanna with a wider range at finer scales. But they respond to other density-based variables with a wider range at coarser scales. The last type of environmental variables has response curves with similar shapes but horizontal shifts across different scales (Figure C-8). It may indicate that African elephants have different optimal

responsive values to these environmental variables at different scales. These variables include Shannon's diversity index, Ratio of savanna, Surface roughness, and Patch richness density. The demonstrated scale dependence may not necessarily mean elephants respond to the environmental variables differently at different scales. It could come from the scaling effects of spatial calculation. Both conditions warn us SDMs are not interchangeable across scales.

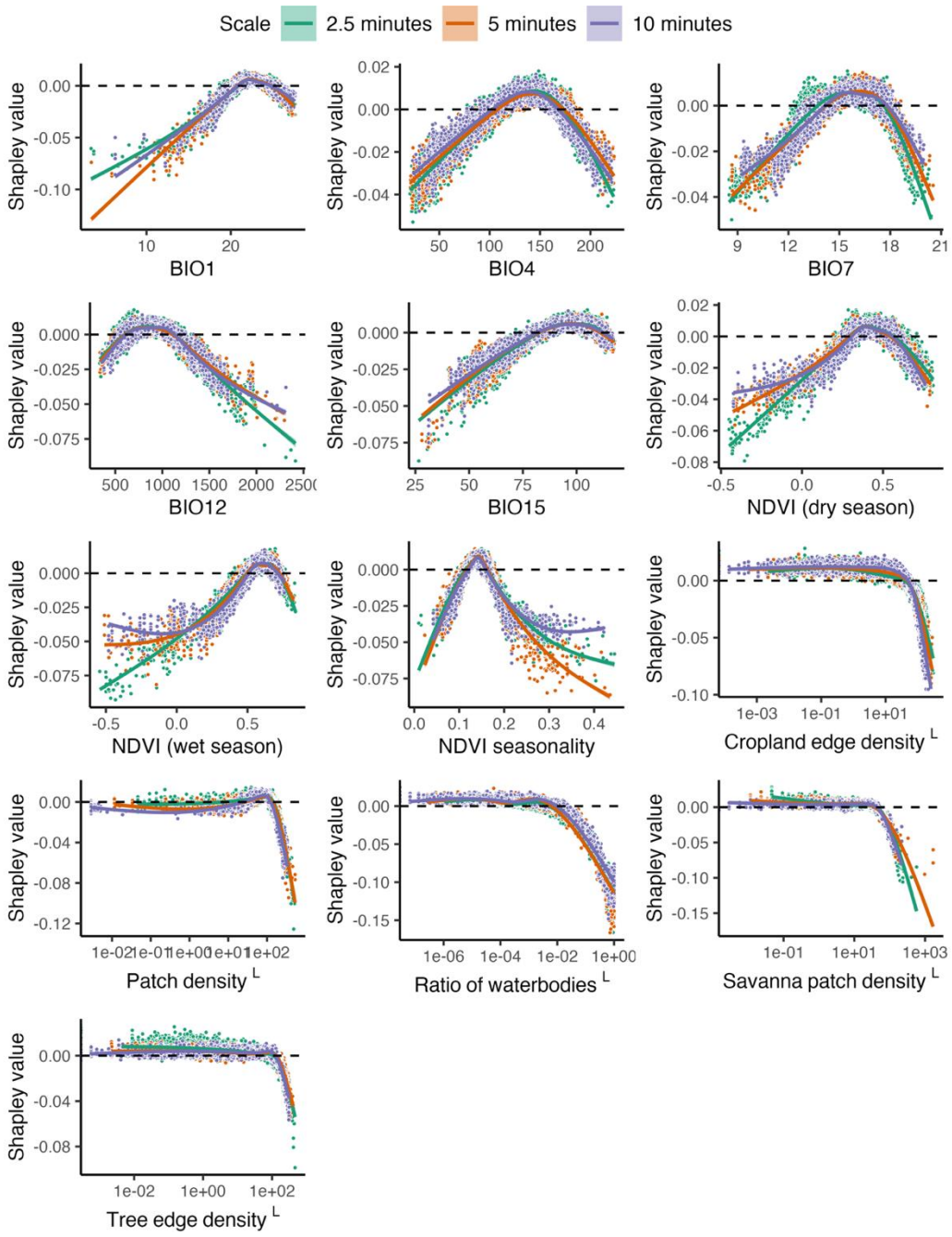

Figure C-6. Statistical responses of African savanna elephants to environmental variables of type A. There is no evident scale dependence for this type. Variables along the x-axis with a superscript ^{L} were $\log_{10}$ transformed. The differences in residual values of seasonal NDVI, NDVI seasonality, and BIOs mainly result from the different grain sizes of doing aggregated calculations (section 3.3.3.1 in the main text).

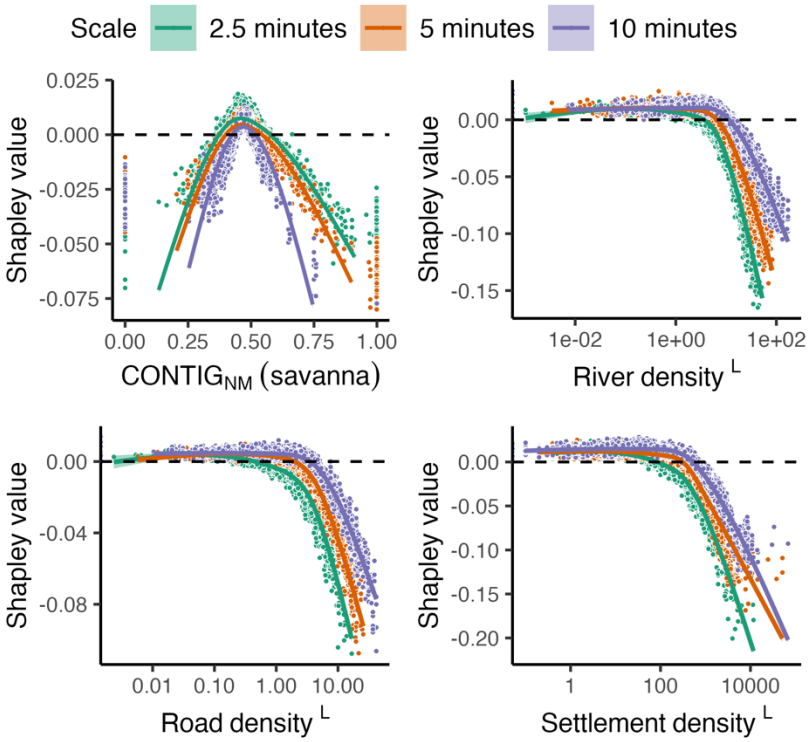

Figure C-7. Statistical responses of African savanna elephants to environmental variables of type B. Response curves have the same central values but different ranges across scales. Variables along the x-axis with a superscript ^{L} were $\log_{10}$ transformed.

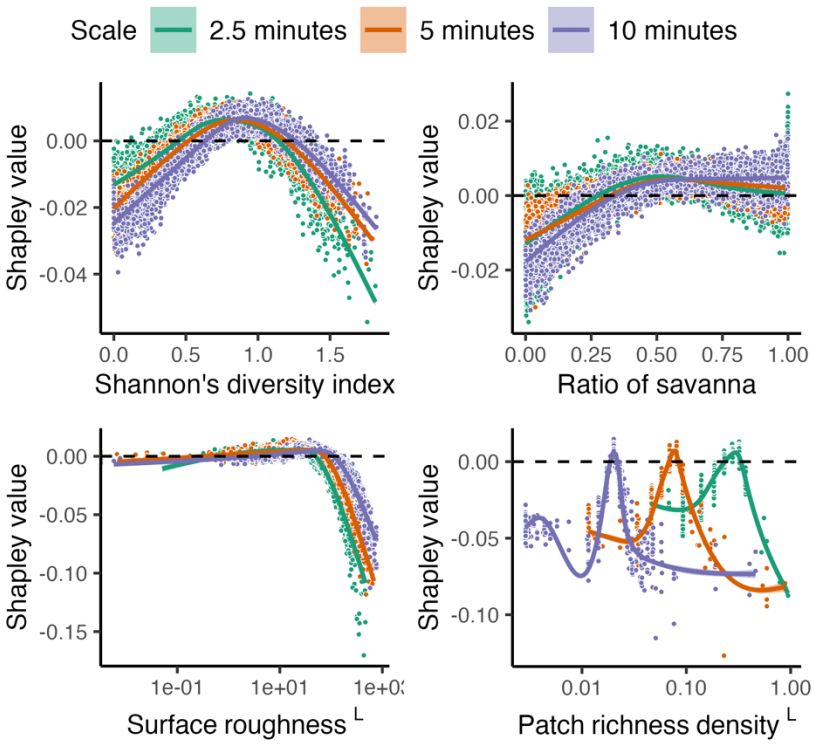

Figure C-8. Statistical responses of African savanna elephants to environmental variables of type C. Response curves have similar shapes but horizontal shifts across different scales. Variables along the x-axis with a superscript ^{L} were $\log_{10}$ transformed.

C.2.3 Supplementary figures and tables

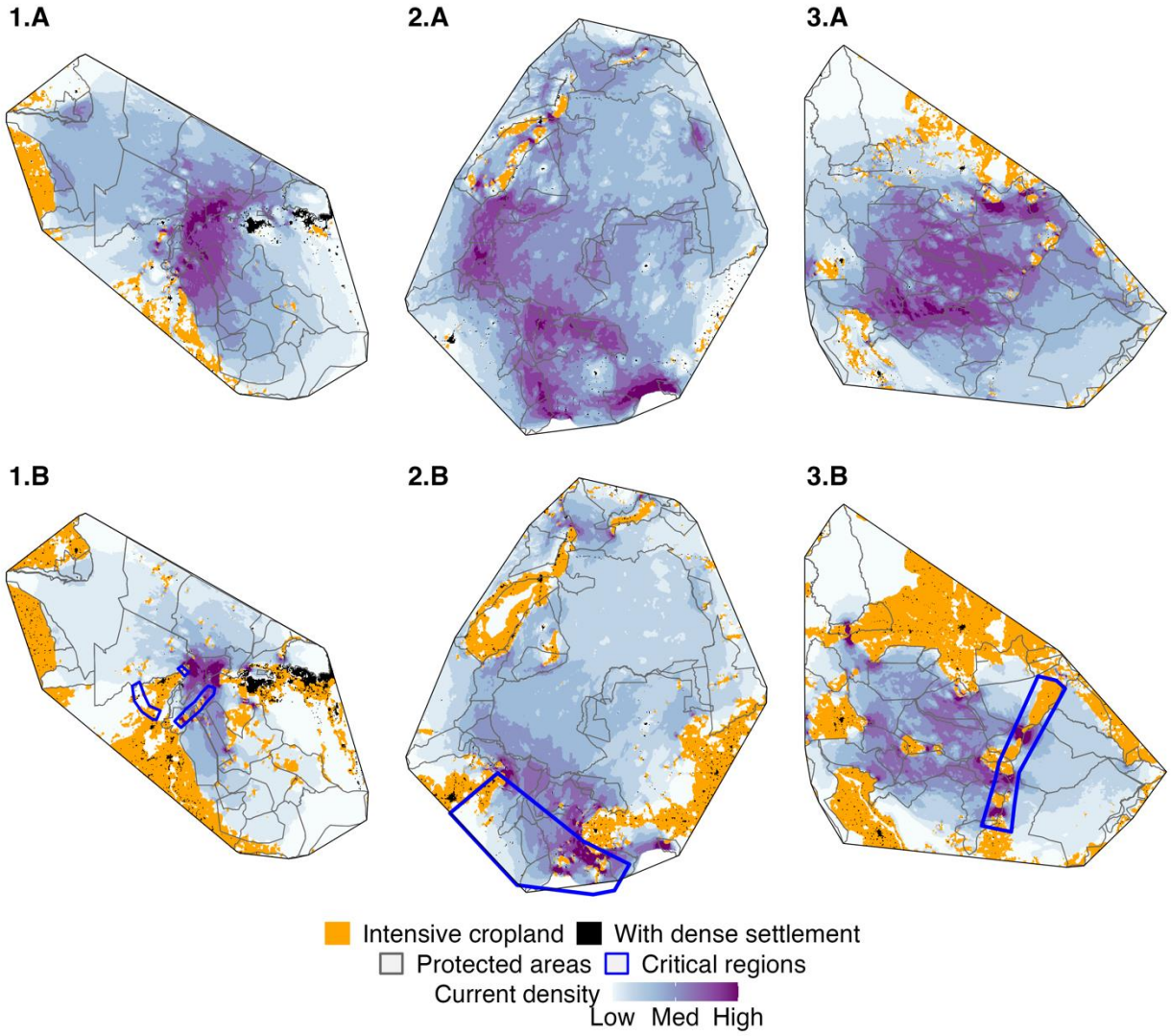

Figure C-9. Landscape connectivity within each primary habitat cluster (No. 1-3 in Figure 3-6 in the main text) under scenario A and B (section 3.3.4) and the detected critical regions of African elephant conservation in Tanzania.

Table C-1. Variables hypothesized to affect African savanna elephant distribution in Tanzania.

Category	Variable	Scale	Original resolution	Description	Data source
Climate	BIO1	Regional only	10, 5, 2.5, 0.5 arc-minutes	Annual Mean Temperature	WorldClim version 2.1
	BIO4	Both		Temperature Seasonality (standard deviation ×100)	WorldClim version 2.1
	BIO7	Regional only		Temperature Annual Range	WorldClim version 2.1
	BIO12	Regional only		Annual Precipitation	WorldClim version 2.1
	BIO15	Regional only		Precipitation Seasonality	WorldClim version 2.1
Vegetation	NDVI (Wet season)	Both	30 meters	Mean NDVI during wet season (Nov - May 2013 - 2019) in each spatial grain at coarse scales or moving windows at fine scale.	Google Earth Engine (Landsat 8)
	NDVI (Dry season)	Regional only		Mean NDVI during dry season (Jun - Oct 2013 - 2019) in each spatial grain at coarse scales.	Google Earth Engine (Landsat 8)
	NDVI seasonality	Both		Mean standard deviation of yearly NDVI (2013 - 2019) in each spatial grain at coarse scales or moving windows at fine scale.	Google Earth Engine (Landsat 8)
Groundwater sources	Coverage of waterbodies	Regional only	4.77 meters	Percentage of deep waterbodies in each spatial grain at coarse scales.	Land cover map made in this study (section C.1.1)
	Rivers	Both	--	Density of rivers (length / area) in each spatial grain at coarse scales or distance to rivers at fine scale.	OpenStreetMap
Antropologic factors	Settlements	Both	--	Density of settlements (count / area) in each spatial grain at coarse scales or distance to settlements at fine scale.	OpenStreetMap
	Cropland edge density	Both	4.77 meters	Edge density of cropland in each spatial grain at coarse scales or moving windows at fine scale.	Land cover map made in this study (section C.1.1)
	Roads	Both	--	Density of primary roads and railways (length / area) in each spatial grain at coarse scales or distance to primary roads and railways at fine scale.	OpenStreetMap
Topographic factors	Surface roughness	Both	30 meters	Standard deviation of elevation in each spatial grain	Advanced Land Observing Satellite (ALOS) DSM
	Vector Ruggedness Measure (VRM)	Landscape only	30 meters	Vector Ruggedness Measure (VRM) in moving windows at fine scale.	Advanced Land Observing Satellite (ALOS) DSM
Landscape metrics (Landscape level)	Patch density (PD)	Regional only	4.77 meters	Patch density for the full landscape mosaic in each spatial grain at coarse scales.	Land cover map made in this study (section C.1.1)
	Patch richness density (PRD)	Regional only		Patch richness density for the full landscape mosaic in each spatial grain at coarse scales.	Land cover map made in this study (section C.1.1)
	Shannon's diversity index (SHDI)	Regional only		Shannon's diversity index for the full landscape mosaic in each spatial grain at coarse scales.	Land cover map made in this study (section C.1.1)
Landscape metrics (Class level)	Coverage of Savanna	Regional only	4.77 meters	Percentage of savanna (shrubland, grassland, and wetland) in each spatial grain at coarse scales.	Land cover map made in this study (section C.1.1)
	Dense tree edge density (ED)	Regional only		Edge density for dense tree patches in each spatial grain at coarse scales.	Land cover map made in this study (section C.1.1)

Savanna contiguity index distribution (CONTIG_MN)	Both	Mean of contiguity index for savanna (shrubland, grassland, and wetland) in each spatial grain at coarse scales or moving windows at fine scale.	Land cover map made in this study (section C.1.1)
Savanna patch density (PD)	Both	Patch density of savanna (shrubland, grassland, and wetland) in each spatial grain at coarse scales or moving windows at fine scale.	Land cover map made in this study (section C.1.1)

All layers at coarse scale were produced by zonal calculation on their original resolution with zones defined by the spatial grain at spatial scales (10, 5, and 2.5 arc-minutes) except the climatic variables. All layers at fine scale (0.5 arc-minutes (~1km)) and optimal scale obtained from analysis at coarse scales were produced by focal calculation on their original resolution with moving window of different sizes (3, 5, and 7 km at fine scale, and the optimal scale at coarse scale) except the climatic variables.

References

Buchhorn, M., Lesiv, M., Tsendbazar, N.-E., Herold, M., Bertels, L., & Smets, B. (2020). Copernicus global land cover layers—Collection 2. *Remote Sensing*, *12*(6), 1044.

Congalton, R., Yadav, K., McDonnell, K., Poehnelt, J., Stevens, B., Gumma, M., Teluguntla, P., & Thenkabail, P. (2017). *Global Food Security-support Analysis Data (GFSAD) Cropland Extent 2015 Validation 30 m V001*.

Cortes, D. (2022). *isotree: Isolation-Based Outlier Detection*. https://CRAN.R-project.org/package=isotree

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, *24*(1), 38–49.

Fourcade, Y. (2016). Comparing species distributions modelled from occurrence data and from expert-based range maps. Implication for predicting range shifts with climate change. *Ecological Informatics*, *36*, 8–14. https://doi.org/10.1016/j.ecoinf.2016.09.002

Hanley, J. A. & others. (1989). Receiver operating characteristic (ROC) methodology: The state of the art. *Crit Rev Diagn Imaging*, *29*(3), 307–335.

Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C., & Guisan, A. (2006). Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, *199*(2), 142–152. https://doi.org/10.1016/j.ecolmodel.2006.05.017

LaGro, J. (1991). Assessing patch shape in landscape mosaics. *Photogrammetric Engineering and Remote Sensing*, *57*(3), 285–293.

Leroy, B., Delsol, R., Hugueny, B., Meynard, C. N., Barhoumi, C., Barbet-Massin, M., & Bellard, C. (2018). Without quality presence–absence data, discrimination metrics such as TSS can be misleading measures of model performance. *Journal of Biogeography*, *45*(9), 1994–2002.

Márcia Barbosa, A., Real, R., Muñoz, A.-R., & Brown, J. A. (2013). New measures for assessing model equilibrium and prediction mismatch in species distribution models. *Diversity and Distributions*, *19*(10), 1333–1338. https://doi.org/10.1111/ddi.12100

Massey Jr, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, *46*(253), 68–78.

McGarigal, K., Cushman, S. A., Ene, E., & others. (2012). FRAGSTATS v4: Spatial pattern analysis program for categorical and continuous maps. *Computer Software Program Produced by the Authors at the University of Massachusetts, Amherst. Http://Www. Umass. Edu/Landeco/Research/Fragstats/Fragstats. Html*, *15*.

Norway's International Climate and Forest Initiative (NICFI). (2020, May 18). NICFI. https://www.nicfi.no/

Ntukey, L. T., Munishi, L. K., Kohi, E., & Treydte, A. C. (2022). Land Use/Cover Change Reduces Elephant Habitat Suitability in the Wami Mbiki–Saadani Wildlife Corridor, Tanzania. *Land*, *11*(2), 307. https://doi.org/10.3390/land11020307

Peterson, A. T., Papeş, M., & Soberón, J. (2008). Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling*, *213*(1), 63–72. https://doi.org/10.1016/j.ecolmodel.2007.11.008

Philipo, M., Ndakidemi, P. A., & Mbega, E. R. (2021). Environmentally stable common bean genotypes for production in different agro-ecological zones of Tanzania. *Heliyon*, *7*(1), e05973.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Song, L., Estes, A. B., & Estes, L. D. (2023). A super-ensemble approach to map land cover types with high resolution over data-sparse African savanna landscapes. *International Journal of Applied Earth Observation and Geoinformation*, *116*, 103152. https://doi.org/10.1016/j.jag.2022.103152

Song, L., & Estes, L. (2023). ITSDM: Isolation forest-based presence-only species distribution modelling and explanation in R. *Methods in Ecology and Evolution*, 2041-210X.14067. https://doi.org/10.1111/2041-210X.14067

Sorensen, T. A. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biol. Skar.*, *5*, 1–34.

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., & others. (2012). GMES Sentinel-1 mission. *Remote Sensing of Environment*, *120*, 9–24.

Turner, M. G., & Gardner, R. H. (2015). *Landscape ecology in theory and practice: Pattern and process* (Second edition). Springer.

Xu, Y., Yu, L., Feng, D., Peng, D., Li, C., Huang, X., Lu, H., & Gong, P. (2019). Comparisons of three recent moderate resolution African land cover datasets: CGLS-LC100, ESA-S2-LC20, and FROM-GLC-Africa30. *International Journal of Remote Sensing*, *40*(16), 6185–6202.

Appendix D
Appendix to Chapter 4

Supplementary Information – Extended methods

Cropland allocation to minimize agriculture-elephant conflict with consideration of biodiversity and carbon costs

D.1 Agricultural yield data

We estimated the current yield of four crops, namely maize, rice, cassava, and pulses, using rainfed crop yields from the newly released global gridded crop dataset in 2015 (GAEZ+ 2015). This dataset was created based on the FAOSTAT database and Global Agro-Ecological Zones (GAEZ) Version 4 global gridded dataset (Grogan et al., 2022). Our decision to use this dataset was based on several factors, namely its high resolution (5 minutes) compared to other crop yield maps (Kim et al., 2021), its incorporation of updated crop data, and its coverage of the crops we are interested in. However, it is worth noting that errors can occur in any agricultural yield estimation dataset, particularly for Sub-Saharan Africa due to the lack of data. The data was downloaded from Harvard Dataverse (https://doi.org/10.7910/DVN/XGGJAV) and cropped to our study area (Figure D-1). To match the planning unit of our analysis (1 km), we downscaled the crop yield maps from 5 minutes to 1 km (Figure D-1) using Random Forest (RF) and relevant variables (see section 4.3.2 in the main text). Table D-1 shows the model accuracy for each crop.

To estimate the attainable yield of maize, rice, cassava, and pulses, we used the water-limited attainable yield maps simulated for the near future (2011-2040) at a 5-minute resolution under RCP 4.5 from GAEZ v4 data. We combined all climate data sources and calculated the average to obtain the attainable yields. For pulses, the maximum attainable yield was considered among chickpea, cow pea, pearl millet, Phaseolus bean, and pigeon pea (Grogan et al., 2022). We downscaled the attainable yield maps to 1 km using a similar method as that used for current crop yield. Figure D-2 shows the original and downscaled maps. However, as vegetation production information was used to downscale the current yield, the downscaled maps of the current yield are more spatially similar to the original maps than the attainable yield maps.

Table D-1. Evaluation of current and attainable crop yield downscaling models

Scenario	Crop	RMSE (t/ha)	R^2
Current	Cassava	1.61	0.45
	Maize	0.17	0.58
	Pulses	0.12	0.26
	Rice	0.55	0.76
Attainable	Cassava	1.18	0.70
	Maize	1.08	0.73
	Pulses	0.43	0.64
	Rice	0.59	0.81

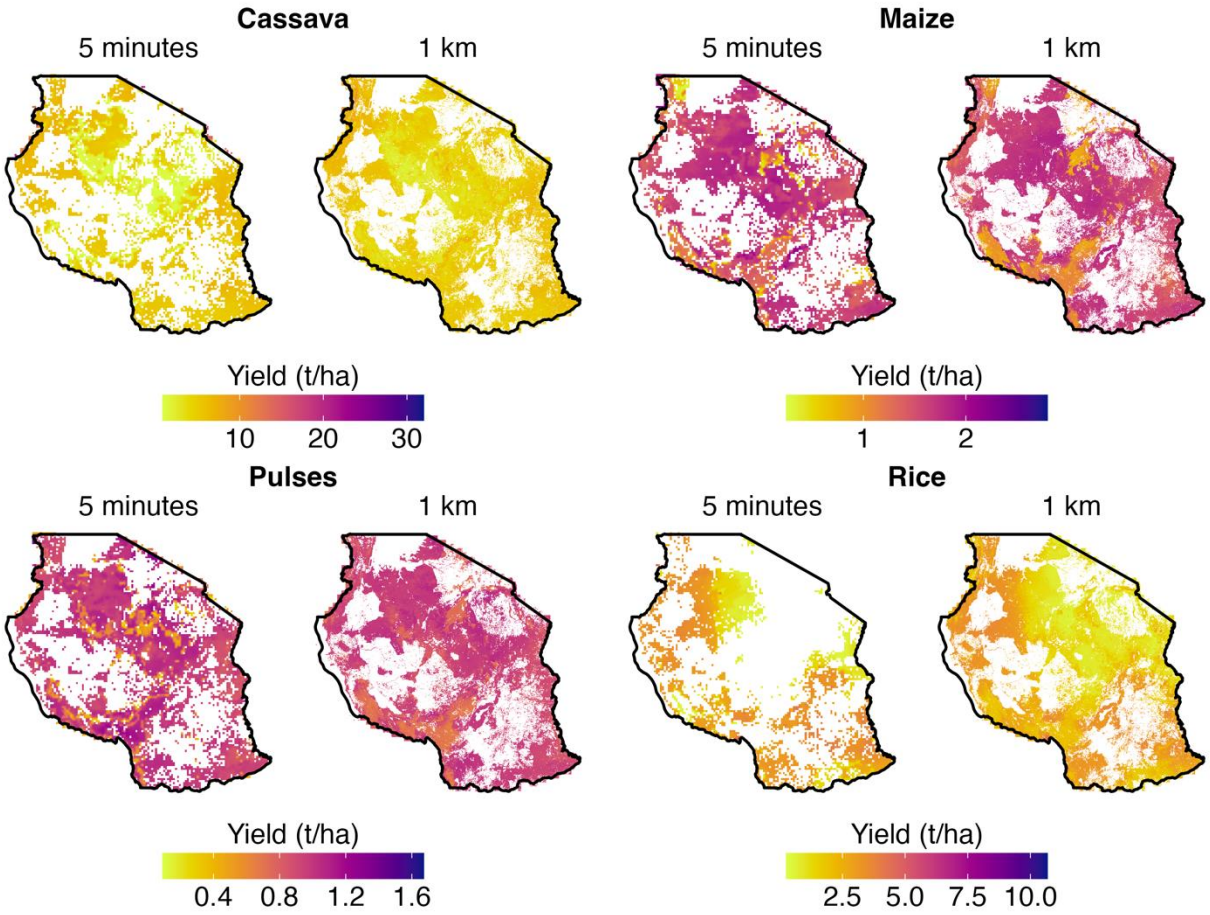

Figure D-1. Current yield map of maize, rice, cassava, and pulses at 5 minutes (A) and the downscaled results at 1 km (B)

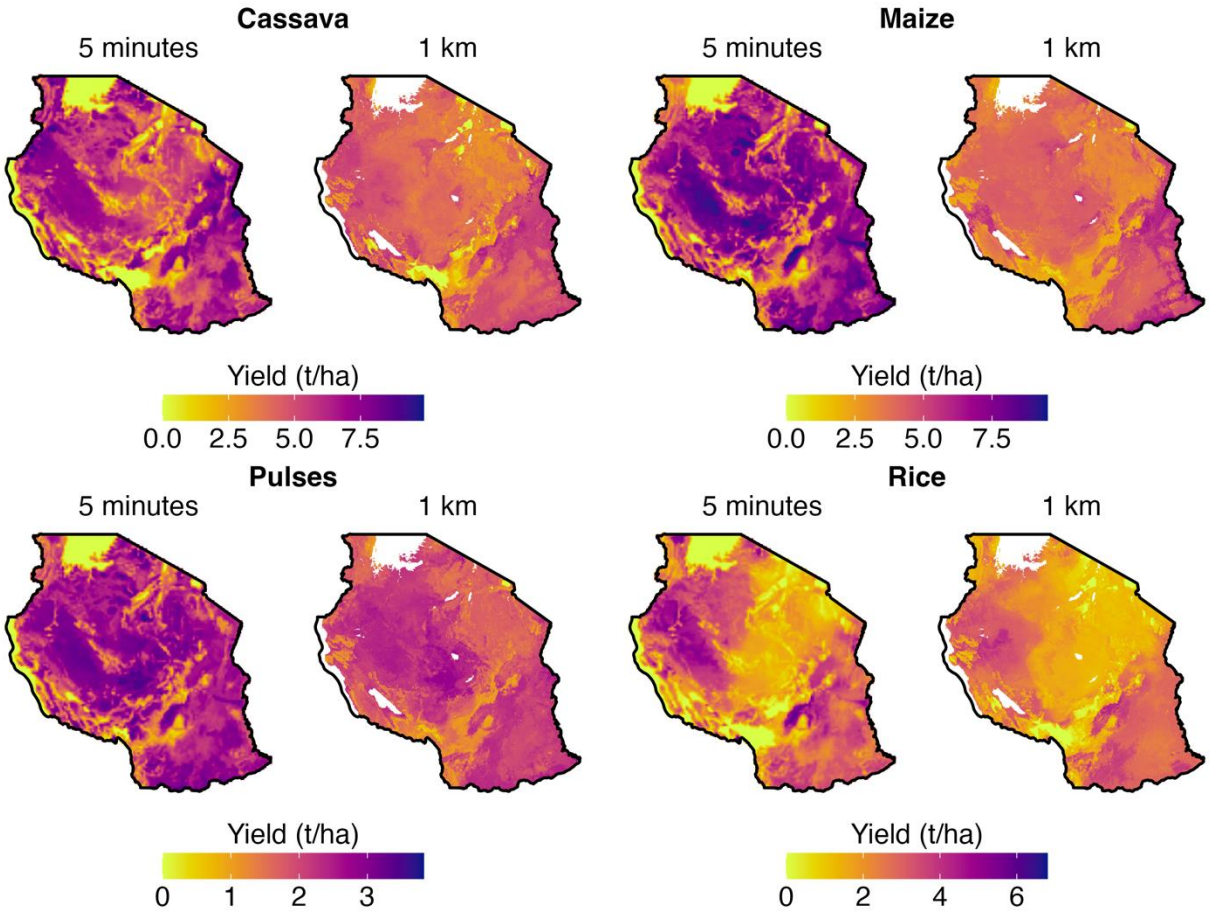

Figure D-2. Attainable yield map of maize, rice, cassava, and pulses at 5 minutes and the downscaled results at 1 km

D.2 Elephant conservation data

Figure D-3 displays the environmental suitability of elephants and census blocks that were created.

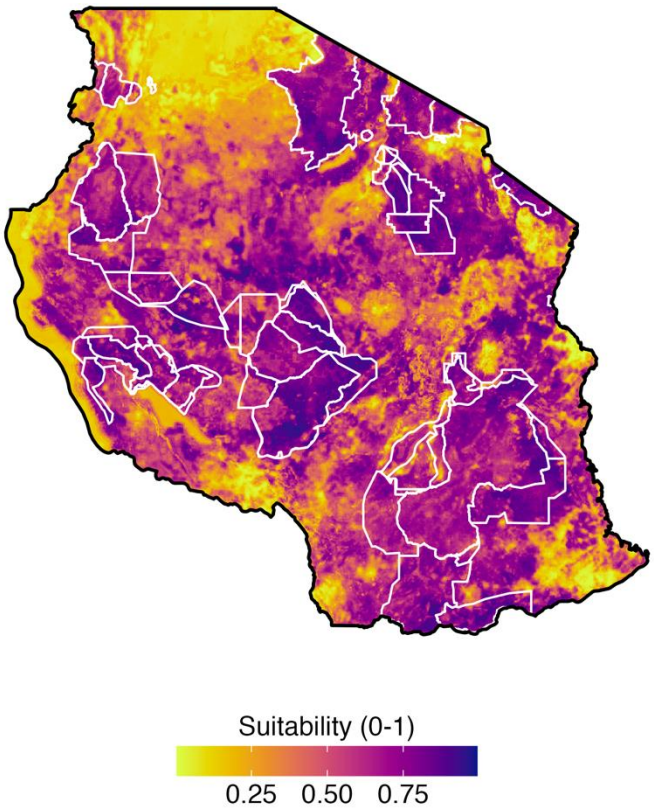

Figure D-3. Environmental suitability of elephants and census blocks (shown in white) used in Circuitscape

D.3 Biodiversity data

As previous studies have demonstrated that using different biodiversity indices in land-use prioritization analyses can produce significantly different results (such as Crawford et al., 2021), we used a proactive index that incorporates multiple perspectives on biodiversity (see section 4.3.4 in the main text). To calculate this index, we compiled species richness and rarity-weighted richness measures based on the geographical ranges of mammals, birds, amphibians, and reptiles. We excluded marine species as well as those classified as Extinct or Extinct in the Wild from the range maps, and only included ranges

where species are considered native, reintroduced, and Assisted Colonization. For migratory birds, we only included ranges where species are considered resident, breeding season, or non-breeding season. We summarized the species from each taxonomic group found in Tanzania in Table D-2.

To reduce overestimation of species richness due to commission errors in the range maps, we refined the range of each species to obtain the area of habitat (AOH) using a global map of terrestrial habitat types (Jung et al., 2020) and Advanced Land Observing Satellite (ALOS) DSM. To further differentiate the contribution of each species to the richness and rarity-weighted richness calculation, we applied multiple weights to species occurrence in each planning unit in a deductive way. Specifically, we assigned full weight to small-range species (range size < global median, Table D-2) and ½ weight to other species. We assigned full weight to endemic species (Table D-2) and ½ weight to other species. Additionally, we assigned different weights based on the IUCN category: CR: 1, EN: ½, VU/DD: ¼, NT: ⅛, LC: 1⁄16. For example, the weight for a vulnerable species with a globally wide range is calculated as: ½ × ½ × ¼. The weighted species richness and rarity-weighted richness measures for the four taxonomic groups are presented in Figure D-4.

The figures in Figure D-5 show the other inputs used to calculate the proactive biodiversity index (Eq. (4-6 and (4-7 in the main text). The figure clearly shows that the Mean Species Abundance (MSA) has a strong correlation with the coverage of protected areas since it considers the relationship between human pressure and impact on biodiversity.

Table D-2. Information of species included in the analysis*

Taxonomic group	No. of species	No. of endemic species	Global median of range size
Mammals	354	29	105622.1 km^2
Birds	1067	29	266659.3 km^2
Amphibians	178	63	2759.8 km^2
Reptiles	347	65	11932.7 km^2

* All the values are calculated from the refined range maps.

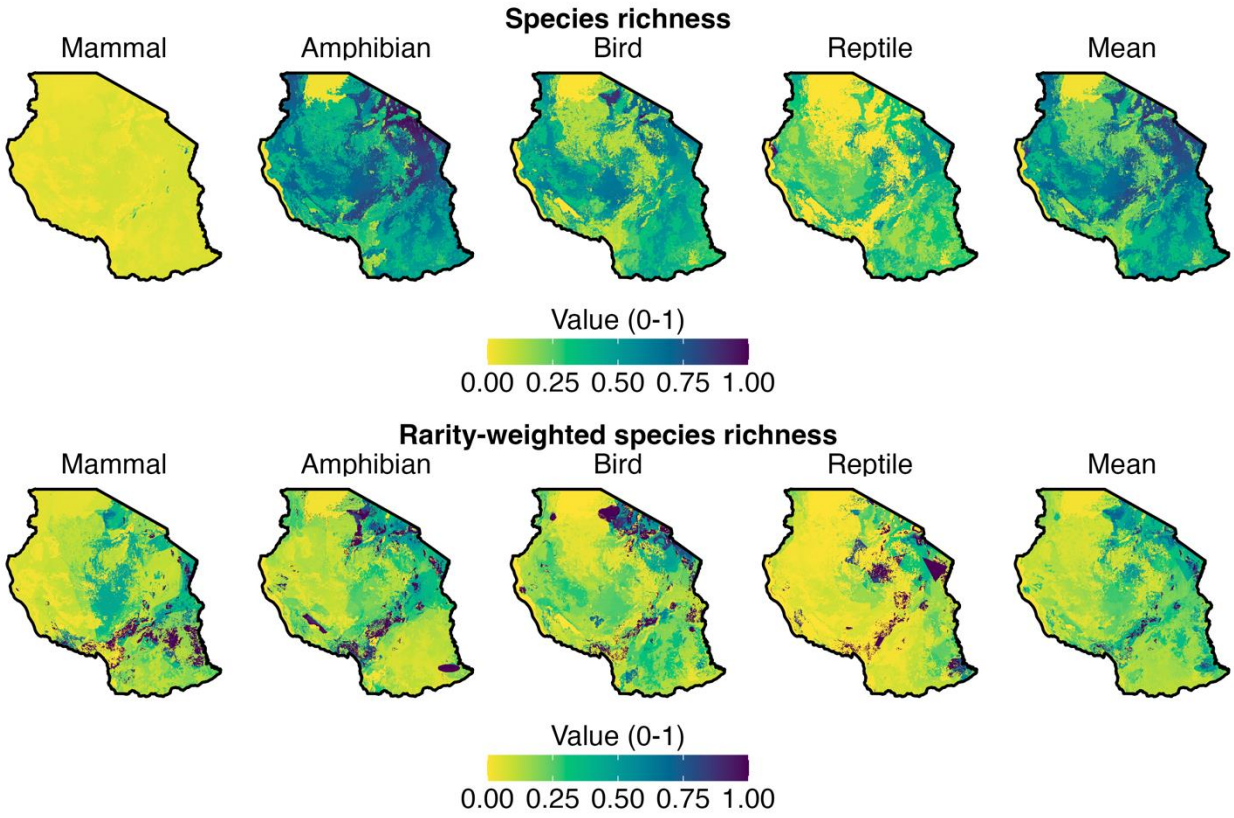

Figure D-4. Weighted species richness and rarity-weighted species richness

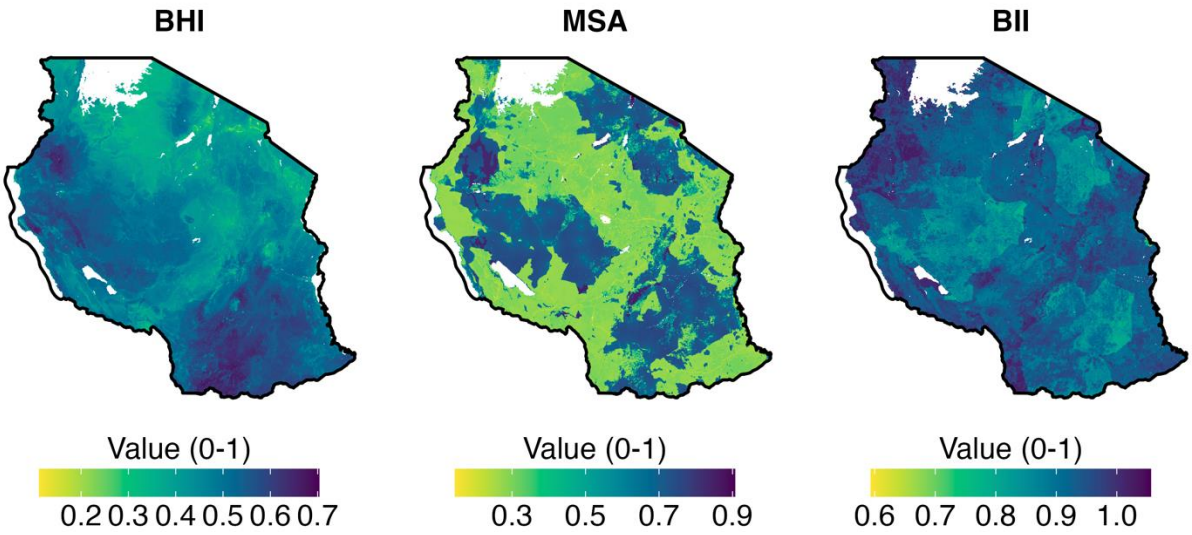

Figure D-5. MSA, BII, and BHI used in the analysis

D.4 Carbon data

The figures in Figure D-6 illustrate the inputs generated for carbon cost calculation (section 4.3.5 in the main text). It is clear from the figure that there is a noticeable spatial correlation between vegetation biomass and soil carbon stocks.

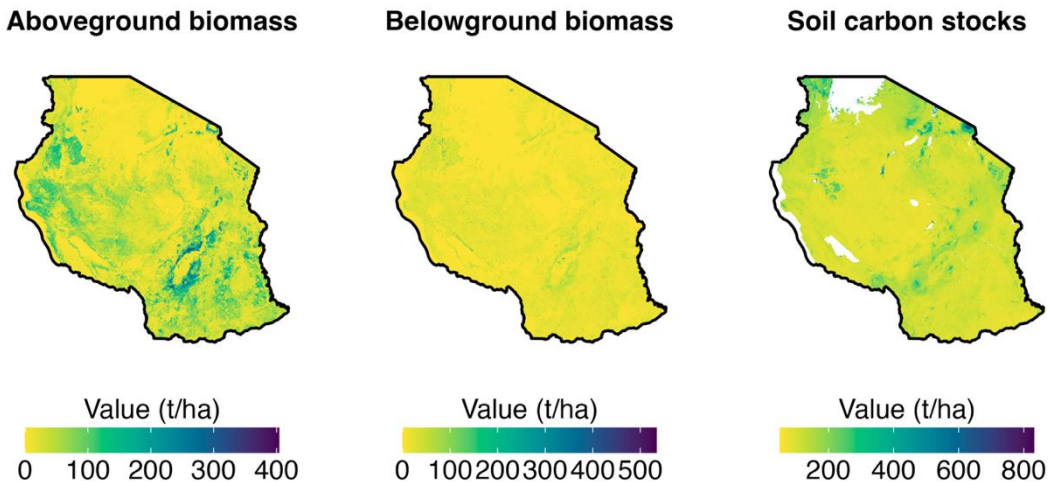

Figure D-6. Aboveground biomass, belowground biomass, and soil carbon stocks in the top 1 m

References

Crawford, C. L., Estes, L. D., Searchinger, T. D., & Wilcove, D. S. (2021). Consequences of underexplored variation in biodiversity indices used for land-use prioritization. *Ecological Applications*, *31*(7). https://doi.org/10.1002/eap.2396

Grogan, D., Frolking, S., Wisser, D., Prusevich, A., & Glidden, S. (2022). Global gridded crop harvested area, production, yield, and monthly physical area data circa 2015. *Scientific Data*, *9*(1), 15. https://doi.org/10.1038/s41597-021-01115-2

Jung, M., Dahal, P. R., Butchart, S. H. M., Donald, P. F., De Lamo, X., Lesiv, M., Kapos, V., Rondinini, C., & Visconti, P. (2020). A global map of terrestrial habitat types. *Scientific Data*, *7*(1), 256. https://doi.org/10.1038/s41597-020-00599-8

Kim, K.-H., Doi, Y., Ramankutty, N., & Iizumi, T. (2021). A review of global gridded cropping system data products. *Environmental Research Letters*, *16*(9), 093005. https://doi.org/10.1088/1748-9326/ac20f4