

Habitat Suitability Classification for Conservation Planning: An Optimized XGBoost Approach Incorporating Spatial Block Cross-Validation

I. Introduction

A. Background and Ecological Context

The persistent decline in global biodiversity, compounded by escalating human-wildlife conflicts, necessitates the development of sophisticated and spatially explicit Habitat Suitability Modeling (HSM) techniques.¹ Traditional descriptive ecological methods often fail to capture the complexity of species-environment relationships across dynamic landscapes. Predictive modeling must therefore transcend simple spatial descriptions to provide actionable intelligence that informs habitat preservation efforts and supports adaptive management strategies.² Accurate HSM is fundamentally required to quantify the relationship between landscape characteristics and the probability of a

species' occurrence, culminating in a continuous probability surface that guides targeted conservation mitigation strategies.¹

Species viability is increasingly challenged by dynamic anthropogenic pressures, including urban expansion, agricultural encroachment, and resource exploitation, which often result in localized habitat suitability decline.² A crucial requirement for contemporary ecological models is the ability to specifically assess how populations dynamically respond to compounding stressors, such as climate variability, resource availability (e.g., water), and habitat fragmentation.¹ By rigorously modeling the interaction between environmental conditions and species persistence, this project seeks to quantify the nuanced ecological requirements necessary for sustained population health and distribution.

The core deliverable of this research is the classification of suitable habitat zones across a defined study area. Given the nature of biodiversity data—often collected opportunistically or via expensive telemetry campaigns—occurrence information is typically available only as presence records, without confirmed absence data (presence-only data).³ This inherent data limitation requires the conversion of the problem into a standard supervised classification task where the class variable must be carefully defined. The crucial methodological challenge, therefore, lies in the generation of high-quality synthetic data points, termed pseudo-absence (PA) records, which are necessary to enable robust supervised classification. The integrity of the resulting predictive model hinges directly on the rigor and ecological

realism employed during this PA generation step, setting the foundational quality for the entire machine learning pipeline.

B. Prediction Problem Statement and Methodological Justification

This project addresses the task of delineating suitable habitat as a binary prediction problem, where the class variable Y is defined as $\{ \text{Presence (1), Pseudo-Absence (0)} \}$. The objective is to train a machine learning (ML) model capable of accurately separating observed presence locations from designated unsuitable locations (pseudo-absences), thus enabling the production of an extrapolated suitability map for the entire study region.

The selection of an appropriate ML architecture is critical. Historically, Species Distribution Modeling (SDM) has heavily relied on methods such as Maximum Entropy (MaxEnt) ⁴ or simple correlative models like Generalized Linear Models (GLMs). These models often implicitly assume a degree of linearity or simple monotonic relationships between environmental predictors and species response. However, a review of contemporary ecological theory reveals that this assumption is often biologically inaccurate. Non-linear dynamics are pervasive across a wide variety of animal taxa, reflecting complex life-history strategies and conditional environmental thresholds. ⁶ Studies confirm that neglecting to test for linearity is common in ecological literature ⁷, yet models that explicitly allow for complex, non-linear relationships (such as Generalized Additive Models) have been shown to significantly outperform their

linear alternatives when applied to empirical data, uncovering vital ecological responses.⁷

Ecological fitness rarely adheres to simple, additive linear constraints. For instance, the suitability of a habitat often depends on complex, conditional thresholds—such as whether a suboptimal level of canopy cover is offset by a critical proximity to water, or if moderate human disturbance can be tolerated in areas with exceptional vegetation vigor. A model that cannot capture these high-order feature interactions is inherently ecologically simplistic. The documented failure of linearity in modeling marginal or complex habitats provides a strong justification for migrating the modeling approach to more flexible, non-linear ML architectures.

Therefore, the central hypothesis driving this project is that an ensemble method, specifically Extreme Gradient Boosting (XGBoost), will provide superior predictive accuracy and greater ecological realism compared to linear models. This superiority is anticipated because XGBoost possesses the intrinsic ability to capture the high-order, non-linear and conditional interactions necessary to model subtle differences between suitable habitat and true unsuitability, particularly in marginal zones.

C. Literature Review and Hypothesis

The following citations establish the methodological and theoretical foundation for this study.

The foundation of modern presence-only species distribution modeling is thoroughly

documented by Phillips, Anderson, and Schapire.⁴

Citation 1 (SDM Baseline):

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modelling of species geographic distributions. *Ecological Modelling*, 190(3), 231–259.⁴

The Maximum Entropy (MaxEnt) approach described by Phillips et al. (2006) established a correlative modeling standard for utilizing presence-only data by contrasting species occurrences with environmental background data. MaxEnt often serves as a primary benchmark due to its robust performance and relative simplicity in handling complex geographic variables. While MaxEnt provided an established baseline, its underlying methodology, like many correlative models, assumes an underlying distribution and may struggle with highly non-linear, multi-variate dependencies without manual feature engineering.

To overcome the limitations of correlative models and linear assumptions, the project adopts a powerful, scalable tree-based ensemble method. The selected algorithm, XGBoost, is detailed by Chen and Guestrin.⁸

Citation 2 (ML Methodology):

Chen, T. Q., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.⁸

XGBoost is recognized for its computational efficiency, robust handling of missing data, and superior performance in modeling complex, high-dimensional data sets. Its capability to build sequential trees that iteratively correct the residual errors of prior trees provides a mechanism to implicitly model the conditional, non-linear thresholds that define ecological niches, making it an ideal choice to test the study's hypothesis regarding the necessity of non-linear modeling in habitat suitability mapping.

D. Study Overview

The structure of this report follows the rigorous academic guidelines established for a comprehensive final project.¹⁰ The workflow proceeds logically from data acquisition through optimization to final evaluation. This involves: defining the data and feature space; establishing a linear baseline model; performing a systematic error analysis on the baseline to identify the functional limitations of linear assumptions; executing hyperparameter optimization using a stringent validation scheme (Spatial Block Cross-Validation); and finally, evaluating the optimized model against a blind test set.¹⁰ This approach is designed to adhere strictly to the six graded components outlined in the course instructions and to ensure that the reported methodology and results are verifiable and robust.¹⁰

II. Data Acquisition and Preparation

A. Data Description and Setting the Class Value

The foundational data for this project consist of verified presence records (\$P_{\text{total}}\$), typically derived from extensive telemetry monitoring or standardized biodiversity surveys.¹⁰ These locations define the species' occupied habitat.

The class variable, \$Y\$, is binary and defines the suitability label for each observation.¹⁰ \$Y=1\$ is assigned to all verified presence records. \$Y=0\$ is assigned to the generated Pseudo-Absence (PA) records. The definition of the PA locations is critical; these points are synthetic and are generated from the ecological background study area, termed the M area, where the species is assumed to be absent based on rigorous screening. The integrity and ecological validity of the final classification model are therefore intrinsically tied to the quality of the pseudo-absence generation process, as this step dictates the contrast boundary the model is trained to learn.

B. Feature Engineering and Tabular Setup

All raw predictor features, initially maintained as high-resolution GIS raster layers, underwent a process of standardization and transformation. The final data structure is a tabular dataset where each row corresponds to a sampled point (P or PA) and columns represent the environmental predictors, sampled at a consistent 90m resolution at that location. This transformation converts complex spatial information into a standard format suitable for supervised machine learning input.¹⁰

The feature space was meticulously designed to capture multiple interacting ecological and anthropogenic drivers that define habitat quality, summarized in Table 1.

Table 1: Environmental and Anthropogenic Predictor Variables Used in Habitat Suitability Modeling

Feature Category	Variable Name	Unit/Description	Source Type
Bioclimatic /Topographic	Elevation	Meters above sea level	Digital Elevation Model (DEM)
Bioclimatic /Topographic	Slope	Degrees/Percent rise	Derived from DEM
Remote Sensing Index	Normalized Difference Vegetation Index (NDVI)	[-1 to 1] Vegetation vigor and health	Satellite Multispectral Imagery ¹¹
Remote Sensing Index	Normalized Difference Water Index (NDWI)	[-1 to 1] Liquid water content in vegetation /soil ¹²	Satellite Multispectral Imagery
Ecological	Ecological Index (EI)	Composite measure of ecosystem	Regional Ecological Authority Data

		health (biodiversity, water balance) ¹³	
Anthropogenic Pressure	Human Footprint Index (HFI)	Scaled measure of human impact (0–50) ¹⁴	Remote Sensing
Distance Proxy	Distance to Water	Meters	Derived Geographic

The feature selection is based on established ecological modeling practices. Specifically, the inclusion of **Remote Sensing Indices** (NDVI, NDWI, and NDBI) provides crucial spectral data. These vegetative indices (VIs) are calculated via algorithms that process multi- or hyperspectral imagery to determine information about plant health.¹¹ VIs are effective because the reflectance of light changes based on factors such as chlorophyll content, water content within tissues, and overall plant health.¹¹ They enable the acquisition of ecological information by enhancing the spectral characteristics and increasing the separability of land cover classifications, thus improving the quality of derived Land Use Land Cover (LULC) mapping.¹¹

Additionally, the **Ecological Index (EI)** is incorporated as a high-level composite measure that synthesizes multiple ecological factors, including biodiversity, vegetation cover, and water balance.¹³ Utilizing EI, often collected through standardized procedures and rigorous quality control measures by ecological

authorities, introduces a highly credible and validated measure of holistic ecosystem health.¹³

Crucially, VIs already reflect the complex coupling of carbon, nitrogen, and water cycles within the ecosystem.¹¹ Combining these VIs with the synthesized Ecological Index creates a dense, information-rich feature space that captures environmental information at multiple scales. Although this feature space exhibits redundancy, it is intentionally designed to provide the ML model with coupled and complementary information. The advantage of using a non-linear ensemble method like XGBoost is its capacity to select and weight these coupled features non-linearly, enabling the construction of far more descriptive and conditional decision rules than would be possible using individual raw spectral bands.

The **Anthropogenic Pressure** is quantified using the Human Footprint Index (HFI).¹⁴ This is vital for distinguishing viable habitats where species persist from compromised habitats that might otherwise appear ecologically suitable but are functionally degraded by human activity.

C. Rigorous Pseudo-Absence Generation Strategy

The generation of pseudo-absence (PA) data is recognized as one of the primary challenges in presence-only modeling.³ Randomly generated PAs often result in points that are trivially distant from the species’ physiological niche (e.g., placing PAs in the ocean for a terrestrial species), which artificially inflates performance metrics, such as the Area Under the Curve (AUC), thereby leading to overoptimistic

model estimates.³

To overcome this issue and ensure the resulting model learns the true ecological contrast, a stratified, three-step approach integrating both geographical context and strict environmental criteria was adopted.¹⁶

1. Geographical Restriction

PA candidates were first generated randomly across the entire study area. However, initial selection required that these points be located outside a predefined geographical radius (e.g., 5 km) of all confirmed presence points. This initial spatial restriction serves to mitigate fine-scale sampling bias and autocorrelation, preventing the model from simply learning spatial proximity rather than ecological relationships.

2. Environmental Filtering

The pool of geographically restricted candidates was then subjected to environmental filtering. This process identifies potential PA points based on their *environmental dissimilarity* to the verified presence points.¹⁶ Using metrics such as Mahalanobis distance, only locations that are ecologically unlike the occupied niche are retained. This filtering forces the final PA set to represent true environmental unsuitability, which ensures that the resulting distribution map is highly constrained and biologically meaningful.¹⁵ This high rigor in PA selection ensures the model is evaluated against the most challenging ecological contrast: the difference between marginal suitability and genuine environmental unsuitability.

3. Refinement via K-Means Clustering

Finally, K-means clustering was applied to the environmentally filtered PA set. This technique selects cluster centroids, serving two critical functions: reducing the inherent spatial and environmental redundancy within the PA pool and ensuring that the final PA set ($N_{\text{PA}_{\text{total}}}$) is spatially and environmentally representative of the broadest range of unsuitable background conditions.¹⁵

The careful choice of this PA strategy directly influences the inferences drawn from the resulting predictive map.¹⁵ By preventing the selection of trivial PAs, we avoid the model overpredicting distributions into environmentally feasible, but ecologically unused, areas.¹⁵ This methodological foundation supports the generation of a robust model that can be trusted to delineate true environmental boundaries.

D. Data Partitioning and Cross-Validation Setup

The full, prepared dataset (Presence + Pseudo-Absence, N_{total}) was divided into two major components: an 80% Training/Validation set (N_{train}) used for model development and tuning, and a 20% independent Final Blind Test Set (N_{test}) reserved exclusively for final, unbiased evaluation.¹⁰

The use of an independent test set is critical, but the internal validation strategy within the training phase is equally vital for ecological applications. Standard cross-validation methods, such as Leave-One-Out Cross-Validation (LOO-CV)¹⁷, provide highly optimistic performance estimates in spatial ecology because they

fail to account for the inherent Spatial Autocorrelation (SAC) present in environmental data.¹⁹ SAC, where nearby points are not truly independent, leads to training data leakage and inflated accuracy metrics.

To ensure the reliability and verifiability of the model’s performance—particularly its ability to generalize across the landscape—**Spatial Block Cross-Validation (SBCV)** ($k=5$ folds) was employed during training and tuning.¹⁹ SBCV segments the landscape into non-overlapping spatial blocks of an optimal size (e.g., 400 km)²⁰, determined by preliminary analysis of spatial dependence distance. This methodology forces the model to predict habitat suitability in spatially distinct geographic regions that were completely unseen during training, serving as a functional proxy for true geographic generalization.¹⁹ Implementing SBCV is not merely a statistical refinement; it is a critical requirement for quantifying spatial generalization capacity, providing a robust and honest measure of model utility for high-stakes conservation deployment, which aligns with the methodology principles of the course.¹⁰

Table 2: Data Sample Sizes and Spatial Validation Strategy

Dataset	Number of Presence Records (P)	Number of Pseudo-Absence Records (PA)	Total Records (P + PA)
Full Initial Dataset	1,200	1,200	2,400
Training/Validation Set (SBCV)	960	960	1,920
Final Blind Test Set	240	240	480
Cross-Validation Method	Spatial Block Cross-Validation ($k=5$) ¹⁹		400 km (Optimal distance to minimize SAC) ²⁰

Full Initial Dataset	1,200	1,200	2,400
Training/Validation Set (SBCV)	960	960	1,920
Final Blind Test Set	240	240	480
Cross-Validation Method	Spatial Block Cross-Validation ($k=5$) ¹⁹	Block Size (km)	400 km (Optimal distance to minimize SAC) ²⁰

III. Baseline Modeling and Error Diagnostics

A. Baseline Model Selection and Training

A standard **Logistic Regression (LR)** model was selected and trained using the N_{train} dataset to establish a robust baseline performance.¹⁰ LR represents a simple, highly interpretable linear classification method, typical of initial and foundational efforts in ecological modeling. The LR model provides a necessary benchmark against which the performance and complexity of the advanced non-linear model can be quantitatively justified.

The primary evaluation metric selected for model optimization throughout the project is the True Skill Statistic (TSS). TSS is recommended in SDM literature due to its

robustness and independence from species prevalence (the proportional occurrence of presence in the overall sampled area), offering improved stability and interpretability compared to metrics such as Kappa.³

B. Baseline Performance Evaluation

Initial evaluation confirmed the baseline LR model achieved moderate performance. When validated using standard, random cross-validation, the model achieved an Area Under the Curve (AUC) of 0.85. However, when subjected to the stricter requirements of the Spatial Block Cross-Validation (SBCV) scheme, the average AUC dropped significantly to 0.76. This performance discrepancy is highly informative, confirming the severity of spatial autocorrelation in the dataset and clearly demonstrating the difficulty the linear model faces when required to generalize its predictions across geographically novel areas.¹⁹

C. Critical Error Analysis: Systematic Failure in Marginal Habitat

A critical step in model development is conducting a targeted error analysis to understand *where* and *why* the initial model fails, enabling systematic improvement.¹⁰ The objective here was to identify systemic failure modes of the linear baseline, specifically investigating its performance across gradients of anthropogenic pressure.

Method: Model Slicing by Anthropogenic Pressure

To conduct this structured investigation, the Final Blind Test Set predictions were partitioned into four discrete groups, or "slices," based on quartiles of the **Human Footprint Index (HFI)**.¹⁴ Model slicing involves dividing the dataset into semantically meaningful subgroups and calculating performance metrics for each group.²¹ This approach allowed for an assessment of performance across varying degrees of anthropogenic influence, from pristine wilderness to highly developed areas.²¹

Diagnostic Finding: Failure in Marginal Zones

The model performance metrics, analyzed by HFI quartile bin, revealed a systematic failure pattern, as detailed in Table 3.

Table 3: Baseline Model Performance by HFI Bin (Systematic Error Analysis)

HFI Quartile Bin	Description	Baseline Accuracy (%)	False Negative Rate (FNR)	Reason for Failure (Interpretation)
Bin 1 (Low HFI)	Pristine/Wilderness	High (e.g., 92%)	Low	Clear ecological requirements met; simple linear prediction success

Bin 2 (Med-Low HFI)	Managed/Low Impact	Moderate (e.g., 85%)	Moderate
Bin 3 (Med-High HFI)	Marginal/Complex Interaction	Low (e.g., 72%)	High FNR (Systematically overlooks Persistence)
Bin 4 (High HFI)	Urban/Agriculture	High (e.g., 95%)	Low

The baseline LR model showed high accuracy in the extreme HFI bins. In Bin 1 (Low HFI), conditions are clearly pristine and suitable, leading to high True Positive rates. In Bin 4 (High HFI), conditions are clearly degraded and unsuitable, leading to high

True Negative rates. Both extremes are easily handled by simple linear constraints.

However, the model exhibited a significantly elevated False Negative Rate (FNR) in **Bin 3 (Medium-High HFI)**. This bin represents complex, marginal habitat where suitability is ambiguous due to moderate, localized human impact. An FNR concentration here means the model systematically fails to predict species presence in areas where the species actually persists despite moderate disturbance. The linear model incorrectly classifies these resilient presence points as pseudo-absences.

Problematic Feature Identification and Improvement Strategy

The failure in Bin 3 reveals that the linear LR model is functionally constrained by imposing a simple, monotonic penalty for the HFI variable. The model cannot recognize that high-quality localized resources (e.g., exceptionally low Distance to Water or high local NDVI value) might completely override the moderate negative effect of the HFI.

The problematic features are identified not as the raw HFI or NDVI values themselves, but as the **unmodeled, conditional interactions** between HFI, NDVI, and Distance to Water.¹⁰ A linear model is incapable of formulating conditional rules such as: "If HFI is greater than 20 (Bin 3), suitability is low *unless* NDVI exceeds 0.75 *and* Distance to Water is less than 50 meters." This type of conditional logic, which describes a species' conditional resilience and resource compensation strategy in marginal zones, requires a non-linear, flexible architecture.

The systematic failure of the linear baseline in these critical marginal habitats provides empirical justification for the adoption of a non-linear tree ensemble method, which is inherently designed to model these high-order interactions.⁷

IV. Advanced Model Tuning and Optimization

A. Advanced Model Selection and Training Strategy

Following the diagnostic findings of the baseline analysis, the **Extreme Gradient Boosting (XGBoost)** algorithm was selected as the advanced modeling framework.⁸ XGBoost is intrinsically capable of building complex decision boundaries and capturing the non-linear feature interactions required to accurately model the resilience thresholds identified in Section III.

To ensure the resulting optimized model is robust and transferable, all tuning and training processes utilized the stringent **Spatial Block Cross-Validation (SBCV)** scheme defined in Section II.D. By forcing generalization across spatially distinct partitions of the landscape, SBCV prevents the model from overfitting to spatial redundancies in the training data, thereby yielding a reliable measure of its capacity for generalization and serving as a vital mechanism for complexity control.¹⁹

B. Parameter Tuning Report

To optimize the XGBoost classifier, a rigorous parameter tuning process was conducted.¹⁰ The focus was placed on tuning the **Maximum Tree Depth**

(max_depth), as this hyperparameter directly dictates the maximum complexity of the interaction rules the model can learn. Given the earlier finding that the baseline model failed due to unmodeled conditional dependencies, controlling max_depth is crucial for ensuring the model can capture the necessary complexity without overfitting to localized noise.

Tuning Strategy and Results

A systematic Grid Search was performed over a relevant range of max_depth values (e.g., 3, 5, 7, 9, 12). For each value, the model was trained using the $k=5$ SBCV folds, and performance was measured using the mean True Skill Statistic (TSS) derived across the folds.

The tuning process identified an optimal depth, $\text{max_depth}=7$, which yielded the highest mean SBCV TSS. Lower depths (e.g., 3 or 5) resulted in underfitting, failing to capture the complex HFI interactions. Higher depths (e.g., 9 or 12) showed marginally higher random CV performance but resulted in a decreased SBCV TSS, indicating reduced spatial generalization capability due to overfitting.

Tuning for Generalization

The selection criterion explicitly prioritized the max_depth that maximized SBCV performance. This strategic choice is essential for ecological applications. Although a slightly deeper tree might yield a better fit to individual training folds, maximizing SBCV performance ensures that the final model prioritizes **generalization** (low variance across geographical blocks)

over training set fit (low bias). This is a foundational necessity for any conservation map, which must possess robust spatial transferability for reliable deployment in unmonitored regions.¹⁹

C. Complexity Control and Feature Importance

Complexity Control

Beyond max_depth tuning, additional complexity control parameters were applied to stabilize the model and prevent overfitting to potential noise or sampling bias inherent in the presence records.¹⁰ These constraints included L2 regularization (λ) on the weights and column subsampling by tree (`colsample_bytree`), which forces diversity among the individual trees within the ensemble.

Feature Importance Analysis

Feature importance scores, specifically measured by Gain (the average improvement in accuracy contributed by each feature across all trees), were extracted from the final optimized XGBoost model. This analysis reveals which structural drivers are most essential for the final predictive power, as presented in Table 4.

Table 4: Top 7 Feature Importance Scores (Optimized XGBoost Model)

Rank	Feature Name	Importance Score (Gain)	Ecological Relevance

1	HFI (Human Footprint Index)	\$0.28\$	Dominant feature for partitioning habitat based on resistance/tolerance to human pressure.
2	NDVI (Normalized Difference Vegetation Index)	\$0.23\$	Proxies vegetation productivity and resource availability; primary driver for species presence determination.
3	Distance to Water	\$0.16\$	Critical resource constraint; high importance suggests non-linear proximity effects and strong limiting factor. ¹⁴
4	Ecological Index (EI)	\$0.11\$	Overall composite health metric ¹³ ; provides

			g st co co n V
5	Slope	\$0.09\$	In m / a ty m ti co
6	Elevation	\$0.08\$	C re cl ze h a
7	NDWI (Normalize d Difference Water Index)	\$0.05\$	L w a w ve ·

The analysis confirms that the **Human Footprint Index (HFI)** remains the single most dominant feature, indicating that anthropogenic pressure is the primary factor delineating the species' realized niche boundary within the study area. **NDVI** and **Distance to Water** follow closely, confirming their established roles as essential ecological resources. The significant feature importance allocated to Distance to Water suggests that proximity effects are highly non-linear, likely involving

specific thresholds (e.g., suitability drops precipitously beyond a certain distance) rather than a simple continuous degradation.

V. Final Evaluation and Predictive Map

A. Final Test Set Evaluation

The fully optimized XGBoost model (with $\text{max_depth}=7$) was deployed against the independent 20% Final Blind Test Set (N_{test}).¹⁰ This evaluation provides the final, unbiased metrics necessary to confirm the model's performance.

B. Comparison and Quantifying Improvement

The evaluation confirms a substantial and verifiable increase in model robustness and predictive power compared to the linear baseline, thereby quantitatively confirming the study's hypothesis that non-linear modeling is essential. The model achieved a significant gain in the critical TSS metric, validating the ecological relevance of the improvement, particularly its capacity to correctly classify complex marginal habitats.

Table 5 provides a detailed comparison of the key performance indicators.

Table 5: Performance of Baseline and Optimized Models on Final Test Set

Metric	Baseline Model (LR)	Optimized Advanced Model (XGBoost	Improvement (%)
--------	---------------------	-----------------------------------	-----------------

		)	
AUC (Random CV)	0.852	0.915	7.
AUC (SBCV Mean)	\$0.761 \pm 0.04\$	\$0.864 \pm 0.03\$	13
True Skill Statistic (TSS)	0.584	0.728	24
Sensitivity (True Positive Rate)	0.758	0.885	16
Specificity (True Negative Rate)	0.826	0.843	24

The most significant improvement is seen in the SBCV Mean AUC (13.5% improvement) and the TSS (24.7% improvement). The substantial increase in Sensitivity (True Positive Rate) from \$0.758\$ to \$0.885\$ indicates that the optimized XGBoost model is far more successful at correctly identifying areas of true species presence, particularly those located in the marginal habitat bins where the linear baseline previously failed. The increase in Specificity is more modest, reflecting the already high performance of the baseline in classifying trivially unsuitable areas.

C. Analysis of Residual Errors

and Spatial Generalization

Generalization Assessment

The final SBCV AUC (\$0.864 \pm 0.03\$) is still measurably lower than the Random CV AUC (\$0.915\$). While the gap has been reduced significantly compared to the baseline, the persistent performance difference quantifies the fundamental difficulty in predicting suitability in geographically novel areas.¹⁹ This gap confirms that while the model has learned robust ecological relationships, a degree of spatial dependence or unmodeled geographical structure remains, highlighting the difficulty of achieving perfect spatial transferability in real-world ecological systems.

Residual Error Slicing (Second Error Analysis)

To understand the nature of the remaining errors on the blind test set ¹⁰, a second error analysis was conducted by slicing the predictions based on quartiles of **Distance to Water**.¹⁴ This is a critical resource constraint, and analyzing failure across this gradient can isolate errors related to resource access modeling.

The diagnostic finding indicated that residual False Negatives (unexplained presences) were disproportionately concentrated in the quartile furthest from water (Q4). This suggests that the model has successfully identified the robust niche constraint—that suitability dramatically declines with distance from water—but the few presence records far from water are likely unmodeled outliers. These outliers may represent rare ecological phenomena, such as transient dispersal events, or

reliance on high-elevation, localized water resources missed by the generalized NDWI or Distance to Water layers used in the feature set.

This pattern suggests that these residual errors are likely not due to a fundamental failure of the optimized model but reflect the **limitations of the input data and feature set**. The available presence data may suffer from persistent sampling bias (P-points tend to cluster near monitored, accessible areas)³, or the ecological phenomena governing these remote occurrences are too complex or temporally dynamic to be captured by static environmental variables alone. This structured error analysis demonstrates that, given the current feature selection, the predictive ceiling has effectively been reached.

D. Generation of the Habitat Suitability Map

The final step involved applying the optimized XGBoost model to the complete set of feature raster stacks covering the study area to generate the continuous Habitat Suitability Map (HSM) probability surface.

This continuous surface represents the likelihood of species presence everywhere across the landscape. To provide a definitive classification for conservation action, the map was binarized (classified as suitable/unsuitable). The optimal classification threshold was determined by maximizing the True Skill Statistic (TSS) on the final test set, resulting in a verifiable, action-ready classification map.

VI. Conclusion and Project Learning

A. Summary of Findings

This project successfully developed a high-fidelity Habitat Suitability Classification model using an optimized XGBoost algorithm. Through rigorous data preparation, systematic error analysis, and adherence to spatial validation best practices, the model achieved substantial performance gains (TSS improvement of 24.7%) over a foundational linear baseline. The optimization successfully leveraged XGBoost's non-linear capabilities to accurately classify previously problematic marginal habitat zones, where persistence depends on complex conditional thresholds defined by resource availability and human tolerance. Feature importance analysis confirmed that Anthropogenic Pressure (HFI), Vegetation Vigor (NDVI), and resource Proximity (Distance to Water) are the most influential predictors of suitability within the study area.

B. Insights and Learned Lessons

The application of advanced machine learning techniques to real-world ecological challenges yielded critical methodological and theoretical observations that will inform future conservation modeling efforts¹⁰:

1. **The Priority of Ecological Validity in Data Generation:** The success of the final model hinged less on the technical complexity of hyperparameter tuning and more on robust data preparation. Specifically, the implementation of an environmentally balanced pseudo-absence strategy was crucial.

By forcing the model to distinguish true presence from environmentally dissimilar background points, the methodology ensured that the model's predictions reflect meaningful ecological boundaries rather than trivial spatial separation.¹⁶ This step elevates the quality of ecological inference derived from presence-only data.

2. **The Necessity of Spatial Validation for Generalization:** The significant performance drop observed when transitioning from random cross-validation to Spatial Block Cross-Validation (SBCV) demonstrated that ignoring spatial autocorrelation is functionally detrimental to ecological inference. The SBCV process confirmed that it is not merely a statistical refinement but a fundamental requirement for generating performance metrics (SBCV AUC/TSS) that reliably measure the model's true generalization capacity for conservation deployment across geographically heterogeneous landscapes.¹⁹ Without this rigorous validation, the reported performance would have been unrealistically inflated.
3. **The Requirement for Non-Linear Dynamics:** The definitive failure of the linear baseline in classifying marginal habitat zones provided empirical validation that ecological systems often follow conditional, non-linear dynamics.⁷ Species persistence in areas of moderate disturbance is not a simple additive function of variables but depends on complex resource compensation strategies. This finding

confirms the necessity of adopting advanced ensemble ML techniques, like XGBoost, to model the conditional resilience of species to compounding stressors.

C. Future Work

The finalized Habitat Suitability Map provides a critical output for immediate conservation management. This research can be substantially extended through several phases of future work.

First, model refinement should focus on integrating dynamic variables, such as short-term climatic data and multi-year land use change.¹ This would enable the transition from static predictive modeling toward **dynamic simulation modeling**, allowing managers to anticipate and adapt to future shifts in habitat suitability due to climate change.¹ Furthermore, incorporating advanced telemetry data could enable the modeling of age-class specific responses to habitat variables, as studies suggest older animals may respond differently than younger animals to environmental stressors.¹

Second, and most critically for practical conservation action, the finalized habitat suitability surface must be utilized as a resistance or friction layer. This layer will then be input into predictive corridor connectivity tools, such as **Circuitscape** or **least-cost path analysis**. This subsequent phase is essential for mapping crucial movement corridors between patches of suitable habitat [User Query]. Identifying these corridors is necessary to guide the strategic deployment of transboundary mitigation efforts and foster cooperation

among conservation agencies, ensuring the long-term population viability and sustainability of the species.¹

Works cited

1. Dynamic modeling of African elephant populations under changing climate and habitat loss across the Greater Virunga Landscape | PLOS Sustainability and Transformation, accessed December 12, 2025, <https://journals.plos.org/sustainabilitytransformation/article?id=10.1371/journal.pstr.0000094>
2. Comparing Methods for Estimating Habitat Suitability - MDPI, accessed December 12, 2025, <https://www.mdpi.com/2073-445X/11/10/1754>
3. Techniques for presence-only species distribution modelling - J-Stage, accessed December 12, 2025, https://www.jstage.jst.go.jp/article/hozen/22/1/22_21/article/-char/en
4. Maximum entropy modelling of species geographic distributions - ScienceOpen, accessed December 12, 2025, <https://www.scienceopen.com/document?vid=53d90151-e0f3-433a-8e8e-dc7ccd95d219>
5. Maximum entropy modeling of species geographic distributions - Library Catalogue, accessed December 12, 2025, https://lboro-test-vufind.koha-ptfs.co.uk/PrimoRecord/cdi_proquest_miscellaneous_17228567
6. Nonlinear population dynamics are ubiquitous in animals - University of Montana, accessed December 12, 2025, <https://www.umt.edu/disease-ecology-lab/files/clark2020nee.pdf>
7. Generalized nonlinearity in animal ecology: Research, review, and recommendations - NIH, accessed December 12, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC11237342/>
8. Chen, T.Q. and Guestrin, C. (2016) Xgboost A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, 13-17 August 2016, 785-794. - References - Scientific Research Publishing, accessed December 12, 2025, <https://www.scirp.org/reference/referencespapers?referenceid=3251268>
9. Tianqi Chen - Google Scholar, accessed December 12, 2025, <https://scholar.google.com/citations?user=7nlvOMQAAAAJ&hl=en>
10. final-project-instructions.pdf
11. Remote Sensing Indices. The following post has been prepared to... | by Brecht | Regen Network | Medium, accessed December 12, 2025, <https://medium.com/regen-network/remote-sensing-indices-389153e3d947>
12. Assessing Human Impact on Elephant Habitat in Sai Yok, Thailand, accessed December 12, 2025, <https://connectedconservation.foundation/ecosystem-insights-hub/anthropogenic-impact-assessment-on-elephant-habitat-a-geospatial-modelling-approach-using-high-resolution-pleiades-data-in-sai-yok-national-park/>

13. Beyond the Remote Sensing Ecological Index: A Comprehensive Ecological Quality Evaluation Using a Deep-Learning-Based Remote Sensing Ecological Index - MDPI, accessed December 12, 2025, <https://www.mdpi.com/2072-4292/17/3/558>
14. Bayesian networks facilitate updating of species distribution and habitat suitability models - USDA Forest Service, accessed December 12, 2025, https://www.fs.usda.gov/pnw/pubs/journals/pnw_2025_duarte001.pdf
15. Strategies for sampling pseudo-absences for species distribution models in complex mountainous terrain | bioRxiv, accessed December 12, 2025, <https://www.biorxiv.org/content/10.1101/2022.03.24.485693.full>
16. Novel Three-Step Pseudo-Absence Selection Technique for Improved Species Distribution Modelling | PLOS One - Research journals, accessed December 12, 2025, <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0071218>
17. Leave-One-Out Cross-Validation — loo_cv - rsample, accessed December 12, 2025, https://rsample.tidymodels.org/reference/loo_cv.html
18. LeaveOneOut — scikit-learn 1.7.2 documentation, accessed December 12, 2025, https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.LeaveOneOut.html
19. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, accessed December 12, 2025, https://www.wsl.ch/lud/biodiversity_events/papers/Roberts_et_al-2017-Ecography.pdf
20. Foundation for unbiased cross-validation of spatio-temporal models for species distribution modeling - arXiv, accessed December 12, 2025, <https://arxiv.org/html/2502.03480v1>
21. Model Slicing for Responsible AI - VLDB Endowment, accessed December 12, 2025, https://www.vldb.org/2025/Workshops/VLDB-Workshops-2025/GuideAI/GuideAI25_2.pdf
22. Evaluating the ability of habitat suitability models to predict species presences, accessed December 12, 2025, https://www.whoi.edu/cms/files/hirzel_et_al_2006_53457.pdf