Advanced Predictive Ecology Framework: Mathematical Modeling and Validation Architecture for the SharkSight Marine Risk System

1. The Foundational Predictive Engine: Binary Logistic Regression (RE)

The initial architecture for the SharkSight system prioritizes operational constraints, deploying a rapidly trained Binary Logistic Regression model to calculate the Encounter Risk Index (RE).[1] This model serves as the hackathon prototype, balancing predictive power against the critical requirement for computational speed and highly interpretable results for public consumption.

1.1 Mathematical Definition and Sigmoid Transformation

The Logistic Regression model is employed to determine the relationship between the environmental features (represented by the input vector X) and the binary probability of shark presence (Y=1).[1] The fundamental output of the model is a continuous probability score, PRisk, which is mathematically constrained to the interval through the Sigmoid function (or Logistic function):
$PRisk=1+e-(\beta 0+\Sigma\beta i\cdot Xi)1$
The calculated PRisk directly represents the estimated Encounter Risk Index (RE).[1] A value approaching unity (1) indicates a high risk of shark presence, whereas a value approaching zero (0) indicates a low risk.[1] The core functional mechanism of the model is the **Linear Predictor**, denoted by ($\beta 0+\Sigma\beta i\cdot Xi$). This term is a weighted linear summation of all input environmental data, known as the log-odds. The coefficients (βi) represent the weights learned by the machine learning process during the brief training period. A coefficient with a large positive value indicates that the corresponding feature (Xi) is a strong predictor of high risk.[1]

1.2 Justification for Logistic Regression Selection

The selection of Logistic Regression for the core SharkSight application is a strategic decision dictated by logistical and communication requirements inherent to the rapid development cycle.

First, this model is one of the most computationally efficient AI models available, requiring minimal processing resources and capable of training in milliseconds or seconds.[1] This efficiency is paramount, meeting the stringent 1-hour time limit established for the rapid AI model training phase.[1] Second, the Sigmoid function is perfectly aligned with the project's classification goal—determining if a habitat is selected (Risk High) or not selected (Risk Low).[1] Its unique ability to transform the complex relationship between multiple factors into a probability score between 0 and 1 is essential for expressing risk intuitively.[1]

Finally, the simple linear combination used in the model guarantees that the resulting prediction is highly interpretable, a critical feature for developing Explainable AI (XAI).[1] Because the relationship is linear, the application can generate dynamic text summaries for the public, directly correlating the risk status to specific environmental factors learned by the model (e.g., attributing high risk to a strongly positive Chlorophyll coefficient (βi), which signifies a preference for high-productivity zones).[1]

This selection of a linear classifier acknowledges that maximum ecological complexity, often captured by non-linear models required for highly migratory species, is momentarily secondary to speed and operational interpretability. This approach represents an operational compromise, ensuring the fastest possible training time for classification to achieve the hackathon's immediate deployment objective.[1]

Table 1: Logistic Regression Formula Components and Interpretation

Component	**Mathematical Name**	**Role in SharkSight Encounter Risk (RE) Prediction**
PRisk	Output Probability	The estimated Encounter Risk Index (0 to 1). High risk → value approaches 1.
e	Euler's Number (≈2.718)	Base for the exponential transformation used in the Sigmoid function.
$\beta 0+\Sigma\beta i\cdot Xi$	The Linear Predictor (Log-Odds)	Weighted sum of environmental data inputs, providing the input to the Sigmoid function.
Xi	Input Features (Vector X)	Satellite-derived environmental variables (SST, Chlorophyll, Bathymetry, SSHA/EKE).
βi	Coefficients (Weights)	Learned parameters determining feature importance and correlation with shark presence (Y=1).

1.3 Risk Index to User Status Conversion

The continuous probabilistic output, PRisk, must be converted into discrete, actionable categories for the user interface, specifically the "Risk Status" Dial.[1]
The defined operational thresholds are:

- **GREEN (LOW Risk):** $RE<0.35$
- **AMBER (MODERATE Risk):** $0.35 \leq RE<0.65$
- **RED (HIGH Risk):** $RE \geq 0.65$ [1]

2. Input Data Vector X: NASA Earth Observation Features and Ecological Linkage

The predictive capability of the Logistic Regression model is entirely dependent upon the input vector X, which comprises features extracted from historical NASA Earth data archives.[1] These features are chosen because they serve as essential oceanographic proxies for the physical and trophic factors known to influence marine predator habitat selection.[1] The model training involves generating a labeled dataset using historical shark detection data (Presence, $Y=1$) and randomly generated Pseudo-Absence points ($Y=0$), paired with the corresponding co-located environmental features.[1]

2.1 Environmental Features and Sourcing

The four primary features used in the final AI training input vector X are detailed below:

1. **Sea Surface Temperature (SST):** Sourced from NASA GIBS API historical data and specified as MODIS L3 Mapped (8-Day) data in the HSI context.[1] Ecologically, SST defines the crucial **thermal niche** of the shark and governs its migratory boundaries.[1]
2. **Chlorophyll-a (Chl-a):** Serving as a PACE Proxy, Chl-a data is sourced from MODIS L3 Mapped (8-Day) products.[1] Chl-a is indispensable as a proxy for **primary productivity**, directly indicating potential foraging areas where zooplankton concentration is high.[1]
3. **Bathymetry (Depth):** Used as a static layer input.[1] Bathymetry provides the necessary **physical constraint** data, helping the model constrain habitat prediction to known preferred depth zones, such as the continental shelf edge.[1]
4. **Sea Surface Height Anomaly (SSHA) / Eddy Kinetic Energy (EKE):** Designated as a SWOT Proxy.[1] These variables are related to mesoscale ocean dynamics, which often aggregate prey resources along fronts and eddies, thereby indicating zones of **dynamic foraging potential**.[1]

Table 2: NASA Environmental Features (Input Vector X) for Habitat Prediction

Feature Variable (X)	Data Source/Proxy	Ecological Rationale for

		Shark Habitat Selection
Sea Surface Temperature (SST)	MODIS L3 Mapped (8-Day)	Defines the thermal niche and limits migratory boundaries (thermal constraints).
Chlorophyll-a (Chl-a)	NASA PACE Proxy / MODIS L3 Mapped (8-Day)	Proxy for primary productivity, identifying potential foraging hotspots and high-productivity zones (trophic dynamics).
Bathymetry (Depth)	Static Layer	Provides a physical constraint, defining habitat preferences such as continental shelf edges or specific depth zones (physical constraint).
Sea Surface Height Anomaly (SSHA) / EKE	SWOT Proxy	Relates to mesoscale oceanographic features that aggregate prey, indicating dynamic foraging potential (ocean dynamics).

2.2 Limitation in Dimensionality and the Need for Augmentation

While the current X vector utilizes essential oceanographic data, it relies fundamentally on 2D surface proxies (SST, Chl-a, SSHA). This presents a known limitation when modeling highly mobile, pelagic predators. Scientific studies confirm that species like the Basking Shark spend significant time, potentially months, at meso- and bathy-pelagic depths, often below the euphotic zone (recorded maximum dive depths exceeding 1,500 m).[1]

Traditional tracking methods (like light-based geolocation) are rendered useless when sharks are below the photic zone.[1] Accurate modeling of these deep-diving species necessitates the integration of 3D data, such as instantaneous depth-temperature profiles.[1] Because the initial Logistic Regression model relies on 2D surface measurements, its output, the Encounter Risk Index (RE), primarily functions as a **surface habitat suitability estimate**. The framework acknowledges that this initial model lacks the necessary 3D spatial context required to capture the full spectrum of vertical habitat utilization, making the planned future upgrade critical for achieving comprehensive ecological fidelity.

3. Advanced Modeling Trajectory: Habitat Suitability Index (HSI) and the Non-Linear Upgrade

The long-term strategy for SharkSight involves transitioning beyond the linear prototype to an advanced, scientifically rigorous system designed for long-term policy relevance. This transition centers on replacing the Logistic Regression model with a Random Forest model to calculate the Habitat Suitability Index (HSI).[1]

3.1 The Random Forest Habitat Suitability Model

The Random Forest model is designated as the 'Prediction Core' for the next-generation system, titled "EcoCast 3D".[1] As an ensemble learning method, Random Forest offers substantial advantages over the linear Logistic Regression by efficiently modeling non-linear interactions and complex high-order correlations among the environmental features (X).[1] This capability is essential for accurately mapping the highly dynamic and complex environmental gradients that drive marine predator movement.[1]

The output of this refined predictive engine is the **Habitat Suitability Index (HSI)**.[1] Like RE, HSI is a single probability score ranging from 0 to 1 calculated for every point on the map, representing the probability of preferred habitat selection.[1] This output is visualized as an HSI Heatmap, transforming raw probability into actionable intelligence—such as informing policy decisions regarding the closure of fishing zones or the issuance of public advisories.[1]

3.2 Scientific Justification: Integrating 3D Movement Ecology

The decision to adopt non-linear modeling and seek 3D validation is a direct response to the documented "safety and management gap" caused by predictive models that rely heavily on historical, surface-biased sightings.[1] The fundamental scientific challenge is that highly migratory species are often untrackable by conventional methods when they occupy meso- and bathy-pelagic depths for extended periods.[1]

The definitive trajectory for scientific validation involves integrating the **Hidden Markov Model (HMMoce) framework**.[1] This framework, which has been shown to provide a 6-fold improvement in error over traditional methods, provides true 3D tracking capabilities.[1]

HMMoce overcomes the limitations of surface-only tracking by integrating **Depth-Temperature Profiles** collected by archival tags. It compares these diagnostic oceanographic signatures against modeled *in situ* oceanographic data from high-resolution products, such as the Hybrid Coordinate Ocean Model (HYCOM, 0.08◦ resolution).[1]

By incorporating this likelihood framework within a state-space model, HMMoce computes posterior probability distributions that estimate the animal's location and, crucially, its **behavioral state** (e.g., "Resident" or "Migratory") at each time point.[1] This transition from relying on 2D surface variables to utilizing 3D physical data and dynamic behavioral metrics is essential for moving the system from a rapid prototype to a definitive predictive ecology platform. The initial Logistic Regression model successfully establishes operational feasibility, and the planned transition to HSI and HMMoce ensures the system addresses the full complexity of 3D movement ecology, securing long-term scientific credibility and policy relevance.

4. Real-Time Validation Architecture: The 'Trophic Sentinel' Conceptual Model

The 'Trophic Sentinel' represents a conceptual, next-generation tagging system designed to provide the critical real-time, ground-truth data necessary for validating current RE predictions and training subsequent, more advanced AI models.[1] This architecture establishes a crucial feedback loop that addresses deficiencies related to data quality and dimensionality inherent in models based solely on historical, satellite-derived inputs.

4.1 Purpose and Functionality

The Trophic Sentinel system is proposed specifically to refine and validate predictions of the Encounter Risk Index (RE) by supplying live behavioral and ecological data.[1] Satellite-only models carry inherent risks of training bias due to reliance on sparse historical tracking data or pseudo-absence points. The Sentinel counters this by confirming biological events and behavioral metrics in real time, transforming static model assumptions into dynamic, ground-truthed observations.

4.2 The Three Sentinel Components

The Trophic Sentinel is conceptually comprised of three integrated modules:

1. **"Motive Monitor"**
   - **Sensor:** Enhanced Accelerometer and Depth Sensor.[1]
   - **Data Contribution:** This module provides real-time measurements of key movement metrics, specifically PSurface (probability of being near the surface) and PVerticalDepth (vertical habitat utilization).[1]
   - **Impact:** The Motive Monitor directly contributes the necessary 3D component missing from the initial NASA satellite feature vector. By quantifying vertical behavior in real time, it allows future versions of the AI model to replace current static assumptions with live, observed animal movement, preparing the model infrastructure for full 3D integration (e.g., HMMoce).[1]
2. **"Water Witness"**
   - **Sensor:** Environmental DNA (eDNA) Micro-Sampler.[1]
   - **Data Contribution:** This provides ground-truthed evidence of trophic dynamics by confirming the presence or absence of prey species.[1]
   - **Impact:** By confirming prey detection, the Water Witness generates highly reliable, high-quality **Presence (Y=1) data points**.[1] This is critical for improving the quality of the training target variable (Y) and mitigating potential bias found in historical training sets, leading to a more robust foundation for the next generation of AI training.[1]
3. **"Instant Alert"**

- **Sensor:** Event-Driven Satellite Link.[1]
- **Data Contribution:** Ensures low-latency communication tied to critical behavioral or environmental events.
- **Impact:** The Instant Alert enables immediate, dynamic updates to the Risk Dial if the Motive Monitor detects an intense behavioral state or the Water Witness confirms a "true foraging event".[1] This ensures the operational application is immediately validated and responsive to critical, real-world events, closing the feedback loop between predicted risk and confirmed activity.

Table 3: The 'Trophic Sentinel' Conceptual Validation System

Sentinel Component	**Sensor/Mechanism**	**Data Contribution to Real-Time RE Refinement**
Motive Monitor	Enhanced Accelerometer & Depth Sensor	Provides real-time PSurface and PVertical Depth measurements, replacing static behavioral assumptions with live 3D metrics.
Water Witness	Environmental DNA (eDNA) Micro-Sampler	Generates ground-truthed Presence points (Y=1) by confirming immediate prey presence, crucial for improving next-generation training sets.
Instant Alert	Event-Driven Satellite Link	Enables immediate, low-latency updates to the calculated Risk Dial based on confirmed critical behavioral events (e.g., active foraging).

5. Synthesis and Trajectory

The SharkSight project presents a mathematically structured, multi-phased approach to marine risk prediction, purposefully navigating the inherent trade-off between computational speed and ecological fidelity.

The foundational layer relies on the **Binary Logistic Regression model** calculating the Encounter Risk Index (RE). This choice is tactical, driven by the need for computational efficiency (training in under one hour) and simple, public-facing explainability, allowing risk attribution directly through the model's coefficients (βi).[1] The input vector X leverages essential NASA Earth Observation proxies, including Sea Surface Temperature, Chlorophyll-a, Bathymetry, and Sea Surface Height Anomaly.[1]

The evolution plan, however, structurally addresses the limitations of this initial approach. The necessity of moving beyond 2D surface proxies—a limitation highlighted by the knowledge that target species are highly migratory deep divers—drives the trajectory toward the non-linear **Random Forest Habitat Suitability Index (HSI) model**.[1] The HSI, which yields a more robust predictive output, is designed to serve a policy function by informing management decisions.[1]

Finally, the conceptual **Trophic Sentinel** validation system is the planned architectural mechanism for data improvement and dimensionality augmentation.[1] The 'Water Witness' component ensures the necessary quality control for training data by providing definitive ground-truthed presence labels.[1] Simultaneously, the 'Motive Monitor' begins the integration of real-time vertical behavior data, providing the initial 3D metrics essential for eventually adopting the full, scientifically validated

Hidden Markov Model (HMMoce) framework.[1]

This multi-stage framework ensures that the system begins as a rapid, functional prototype while maintaining a clear, scientifically validated trajectory toward a policy-relevant 3D predictive ecology tool capable of accurately modeling complex pelagic habitat selection.

Works cited

1. RESEARCH.pdf