

---

# A Data-Driven approach to Detecting Structural Mispricing in Tybee Island Properties

---

**Prasun Banerjee**

Department of Computer Science  
Georgia Institute of Technology  
Atlanta, GA 30332  
pbanerjee34@gatech.edu

**Ananya Shetty**

Department of Computer Science  
Georgia Institute of Technology  
Atlanta, GA 30332  
ashetty89@gatech.edu

## Abstract

Recent literature suggests that the American coastal real estate market contains a "climate bubble" characterized by billions of dollars in overvaluation due to unpriced flood risks. In this study, we investigate this phenomenon by conducting a granular analysis of property values on Tybee Island, Georgia. We leverage high-resolution elevation data water level time series data streamed from SEC-OORA hydrological sensors to quantify property-specific flood exposure. Our methodology integrates these datasets to engineer novel risk metrics, including hydrostatic probability of exceedance and site-specific freeboard calculations, which serve as inputs for a gradient-boosted regression framework. By employing Quantile Regression, we construct a Rational Price Corridor to distinguish between structural value and speculative pricing errors driven by ignored environmental hazards. We suspect that the structural overvaluation we find on Tybee Island may be the result of beach-front view or other location-based premiums outweighing climate-risk discounts. Our contributions include lightweight Data-Loader software to efficiently scrape and structure historical water-level records for analysis, a batch streaming ETL pipeline to power real-time inference, along with rigorous statistical analysis quantifying and visualizing property mispricing on Tybee Island. While our analysis is restricted to the Tybee domain, we demonstrate a scalable framework for detecting climate-risk overvaluation that is potentially generalizable to other high-risk coastal zones.

## 1 Background

The stability of the American real estate market is increasingly threatened by a "climate bubble" driven by the systemic mispricing of environmental hazards. Recent empirical research suggests that coastal residential properties are overvalued by hundreds of billions of dollars due to unpriced flood risk alone. This market failure persists because current prices often fail to reflect the growing probability of inundation, fueled by information asymmetries, reliance on outdated federal flood maps, and cognitive dissonance regarding the severity of future climate scenarios.

This information gap has been exacerbated by actions such as Zillow's decision in 2025 to remove detailed climate risk scores from individual property listings following complaints from real estate groups. This lack of transparency reinforces existing market inefficiencies, making it increasingly difficult for buyers to assess chronic flood risk before purchase. Consequently, many coastal communities face a looming correction where asset values could precipitously decline as risk perceptions align with hydrological reality.

This study explores the influence of flood risk on home prices within a specific coastal market: Tybee Island, Georgia. As a low-lying barrier island, Tybee faces a prominent and increasing risk from tidal flooding and storm surges. Despite this physical vulnerability, the island maintains strong demand

and high property values, suggesting that buyers may be continuing to pay premiums that outweigh future flooding concerns. This makes Tybee Island an ideal test case for clarifying how climate risk and market prices interact at a granular, local level.

We aim to determine whether this localized market efficiently capitalizes chronic flood exposure into property valuations. Specifically, we address this problem through the following methodological steps.

1. **Data Ingestion and Structuring:** We developed an automated framework utilizing Apache Airflow and custom Python scripts to integrate high-frequency water level data from SECOORA IoT-enabled hydrological sensors on Tybee Island. This process converts raw RESTful API data into a clean, structured historical time series.
2. **Feature Engineering:** Leveraging the corrected historical water level data, we constructed a suite of dynamic, parcel-specific flood risk signals that integrate static property elevation with annual hydrological statistics. Through Lasso regression, we identify the most salient risk metric, finding that duration-based signals (e.g., the frequency of near-miss events) are the strongest predictors of price variance.
3. **Hedonic Pricing and Residual Analysis:** We utilize a non-linear XGBoost hedonic regression model to estimate a "baseline" property valuation based only on standard structural and locational controls. We then perform statistical hypothesis testing on the resulting price residuals, regressing them against the isolated flood risk signals.

By linking precise, data-driven flood risk signals to the residuals of the baseline price model, this study tests the null hypothesis that flood risk is uncorrelated with the unexplained variation in property prices.

## 2 Literature Review

The capitalization of environmental hazards into real estate markets has become a central focus of climate economics, particularly as the frequency of extreme weather events accelerates. A growing body of literature indicates that the U.S. housing market systematically underprices flood risk, creating a climate-risk induced bubble that threatens the financial stability of homeowners and municipalities. Gourevitch et al. (2023) found that residential properties in the United States are overvalued by approximately \$121 billion to \$237 billion due to unpriced flood risk alone. This finding is corroborated by broader market analyses; for instance, a 2025 report by Realtor.com estimated that over \$12.7 trillion in real estate assets face severe climate risks, much of which remains opaque to buyers due to outdated federal maps and voluntary disclosure laws. Similarly, Hino and Burke (2021) demonstrated that single-family homes in floodplains are overvalued by nearly \$44 billion, noting that "markets fail to fully account for information about flood risk," particularly in communities with lower risk awareness. The persistence of this pricing inefficiency is largely by information asymmetries. Buyers often lack access to granular, forward-looking risk data, forcing them to rely on binary FEMA designations that fail to capture the nuance of hydrostatic risk or future sea-level rise scenarios. This informational gap was recently exacerbated in December 2025, when Zillow, a leading real estate marketplace, removed climate risk scores from its property listings. As reported by The Guardian, this decision was made following complaints from industry stakeholders that the data were "hurting sales," effectively removing a critical layer of transparency for prospective buyers and further forcing the information asymmetry that fuels overvaluation. Although some studies indicate that sophisticated commercial buyers are beginning to discount flood-prone assets, the residential market remains largely sticky. Research by Milliman (2022) highlights that only 4% of U.S. homeowners carry flood insurance, reflecting a widespread misconception that lack of a mandatory requirement equates to lack of risk. Our study builds on this literature by moving beyond binary risk classifications. By integrating elevation data with site-specific hydrological modeling for Tybee Island, we aim to statistically model this rampant mispricing by regressing the residuals of a baseline fair-value model calibrated exclusively on standard structural attributes against a suite of non-binary, complex risk signals. We demonstrate that these risk metrics fail to provide statistically significant explanatory power for the price residuals, empirically demonstrating that granular flood exposure is not currently capitalized into Tybee Island asset values, confirming the existence of systemic market inefficiency.

### 3 Data Infrastructure and Pre-Processing

#### 3.1 Data-Loader Software

The foundation of our analysis relies on high-frequency water level data retrieved from the Burton 4H sensor, strategically located in the center of Tybee Island. This sensor streams real-time environmental observations to the Southeast Coastal Ocean Observing Regional Association (SEC-OORA) network. Access to this historical archive is provided via a RESTful API endpoint (<https://api.sealevelsensors.org/v1.0>), which exposes the data in a semi-structured JSON format. A single observation record arrives in the following raw format, containing metadata regarding the gateway signal strength (`rsssi`, `snr`) and timestamping (`phenomenonTime`), alongside the primary scalar value (`result`).

```
1 {
2   "@iot.selfLink": "https://api.sealevelsensors.org/v1.0/Observations(27099756)",
3   "@iot.id": 27099756,
4   "phenomenonTime": "2025-06-20T23:06:29.345927Z",
5   "resultTime": "2025-06-20T23:06:29.345927Z",
6   "result": -2.924,
7   "parameters": {
8     "gateways": [
9       {
10        "snr": 8,
11        "name": "gatech-coastal-108",
12        "rsssi": -98
13      }
14    ],
15    "gateway_count": 1
16  }
17 }
```

Listing 1: Sample Sensor JSON Output

While the API follows OGC SensorThings API standards, the sequential nature of pagination presents a significant bottleneck when retrieving years of high-resolution historical data. To address this latency and create a unified, structured dataset, we developed a custom Python-based ingestion engine (see `DataLoader.py` in the released Github source code). The core of this extraction pipeline is the `DataLoader` class, which implements a concurrency-based fetching strategy. We utilized the `concurrent.futures.ThreadPoolExecutor` library to parallelize the data retrieval process. The `extract_data_concurrent` method decomposes the target historical time window into smaller, non-overlapping chronological chunks (by month or year). These chunks are processed simultaneously across multiple worker threads, bypassing the throughput limitations of a single-threaded sequential fetch.

The raw JSON responses are parsed and flattened into a pandas `DataFrame`, extracting only the essential `phenomenonTime` and `result` fields while discarding extraneous gateway metadata. The system ensures data integrity by enforcing UTC timezone standardization and removing invalid or missing entries. Finally, the aggregated data is sorted chronologically and exported to the CSV format for downstream feature engineering and modeling. This efficient software architecture allowed us to reconstruct a complete historical timeline of water levels from the sensor with minimal computational overhead.

#### 3.2 ETL Pipeline

While the retrospective analysis presented in this study relies on the static historical dataset described in Section 3.1, establishing a robust mechanism for continuous data ingestion is critical for maintaining the informational relevance of the flood-risk signals we discuss in Section 4.1. To facilitate this, we architected a production-grade Extract, Transform, Load (ETL) pipeline orchestrated by Apache Airflow. The system is designed to batch-stream high-frequency water level readings from the Tybee Island `gt-envsense-069` IoT sensor (Datastream 262) and archive them for longitudinal analysis. The workflow is defined as a Directed Acyclic Graph (DAG) named `sensor_pipeline`, which enforces a modular execution flow consisting of four primary components.

1. `create_table()` ensures the persistence layer is prepared by verifying the existence of the target PostgreSQL table and validating the schema.

```

1 CREATE TABLE IF NOT EXISTS water_data (
2     timestamp TIMESTAMP NOT NULL,
3     result NUMERIC(10, 4),
4     PRIMARY KEY (timestamp)
5 );
6

```

Listing 2: Sample Sensor JSON Output

2. `extract_sensor_data()` ingests the latest window of water level observations.
3. `transform_sensor_data()` processes the raw JSON response—stripping nested gateway metadata and parsing the OGC SensorThings format into a clean, structured tabular dataset.
4. `load_data_to_postgres()` commits the transformed data to a PostgreSQL database.

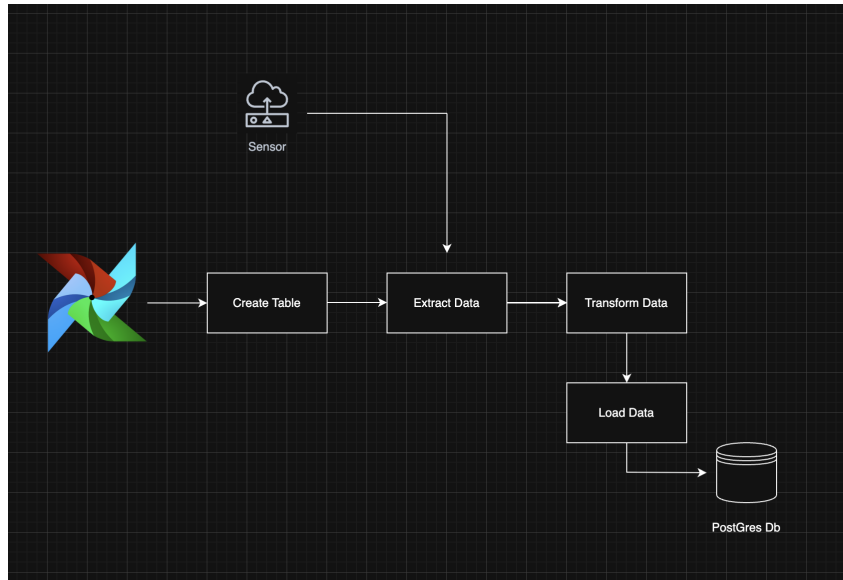


Figure 1: ETL Pipeline Architecture

### 3.3 Hydrological Data Pre-processing

Before engineering flood risk features, it was necessary to subject the raw telemetry data to a rigorous cleaning and harmonization process to ensure longitudinal consistency. Initial quality control involved the removal of transient sensor noise. We applied a range-based filter to exclude physically implausible outliers, strictly retaining observations within a trusted interval (defined as  $\pm 5$  units relative to the mean water level). This step eliminated extreme values attributable to instrument malfunction or transmission errors that could otherwise skew the probability distributions used in our risk modeling. Such extreme spikes in the time-series are often attributable to transient physical obstructions, such as floating logs or debris impacting the sensor apparatus, as well as anthropogenic disturbances including vessel wakes, sensor malfunctions, or localized electrical noise within the telemetry system.

Following outlier removal, a visual inspection of the time series revealed a significant structural break in the data. This discontinuity was characterized by a permanent shift in the mean water level, attributable to a documented change in the sensor's physical location and elevation during the recording period. Leaving this break uncorrected would have introduced a systematic bias, artificially inflating or deflating the flood risk estimates for the period following the sensor relocation.

To rectify this, we implemented a programmatic detection and correction routine. We defined a specific search window (indices 190,000 to 204,000) where the shift was visually observed. Within this window, we identified the precise changepoint by locating the minimum first-order difference in the time series, which corresponded to the moment of the sensor displacement. We then calculated

the local mean water levels for the periods immediately preceding and following the break point, incorporating a buffer to ensure that the means were not influenced by immediate transition noise. The difference between these two means yielded a shift value, which was applied to the subsequent data points (see Figure 3) This bias correction harmonized the time series, effectively normalizing the entire dataset to a single vertical datum and ensuring that the water level distribution remained consistent over the full historical period.

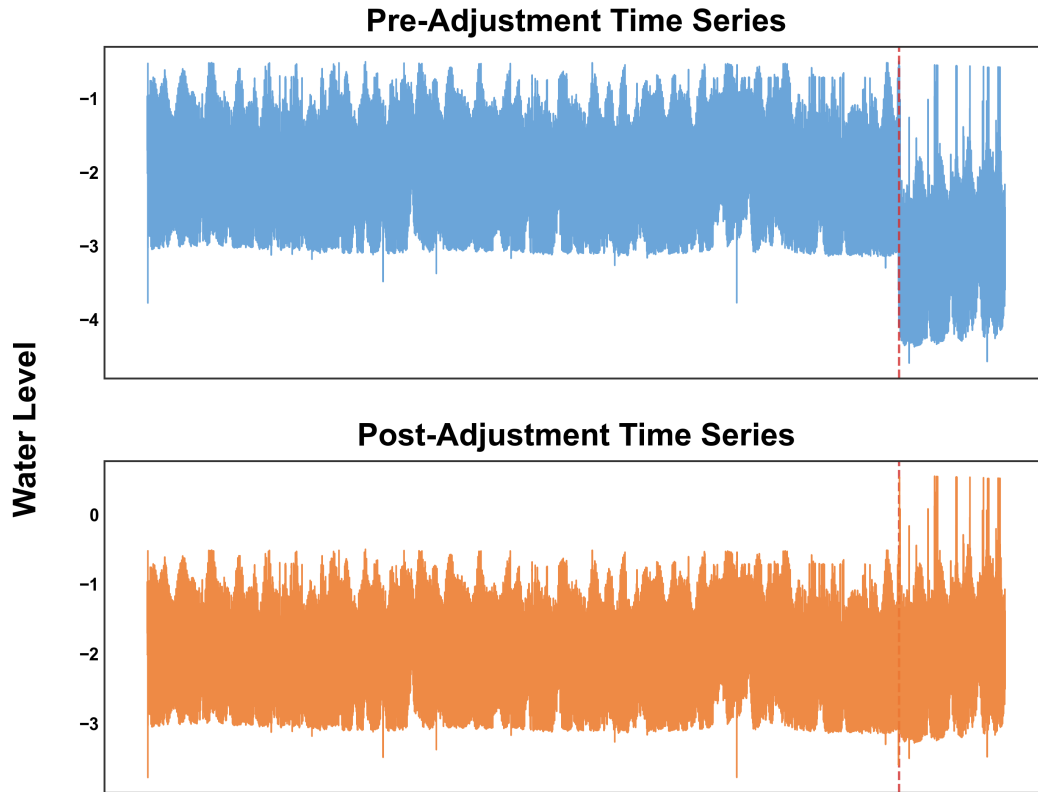


Figure 2: Time-Series Before and After programmatic adjustment

### 3.4 Parcel Data Preprocessing

To construct a robust multi-year dataset for Tybee Island, we implemented a standardized preprocessing pipeline to harmonize parcel files from 2020 through 2023. This process ensured uniform variable definitions across years where naming conventions varied. We reduced the raw schema to a focused set of valuation drivers, retaining essential structural features such as Year Built and Acres, location variables including Neighborhood Code and Land Use, and transaction-level data such as Sale Price and Sale Date. Redundant value metrics were pruned; for instance, since Total Assessment is a linear proxy for Fair Market Value, we retained a single metric to avoid multicollinearity. Static geographic variables like State and Municipality were removed as they offered no variance within the Tybee Island micro-market.

Strict filtering logic was applied to ensure economic validity. We removed non-qualified sales and entries containing null values or zero-valued indicators for price, assessment, or acreage, as these do not reflect arms-length market transactions. Feature engineering included the derivation of property Age at the time of sale and the application of logarithmic transformations to Sale Price and Acres to stabilize variance. Categorical fields were formally encoded, and a custom Flood Index was integrated based on our hydrological forecasts. The final output was a unified, consistent panel dataset ready for regression modeling.

## 4 Modeling Flood-Risk

### 4.1 Defining Flood-Risk Signals

To quantify the exposure of individual properties to tidal flooding, we engineered a suite of dynamic risk signals that integrate static property characteristics with high-frequency temporal water level data. The core of this process involves mapping the static elevation of each land parcel ( $E$ ) to the distribution of water levels ( $w$ ) observed in the corresponding year. The feature engineering process iterates through the dataset on an annual basis. For each year  $y$ , we extract the full distribution of water level observations from the Tybee Island sensor. We calculate aggregate statistics for the water year including the maximum ( $w_{max}$ ), mean ( $w_{mean}$ ), and standard deviation ( $w_{std}$ ) as well as key percentiles ( $w_{p90}$ ,  $w_{p99}$ ). These hydrological metrics are then compared against the elevation of every property to generate parcel-specific risk metrics. Table 1 provides a formulaic description of our features.

### 4.2 Feature Importance

To rigorously isolate the most informative flood risk signals from our engineered feature set, we employed a Lasso (Least Absolute Shrinkage and Selection Operator) regression framework. This approach was necessitated by the complex structure of our feature space and the specific statistical challenges posed by high-dimensional hydrological data. A preliminary analysis of our engineered feature space revealed extreme multicollinearity among the flood risk signals. As illustrated in the correlation matrix (Figure ??), clusters of derived variables exhibit correlation coefficients approaching 1.0. In standard Ordinary Least Squares (OLS) regression, this multicollinearity renders the model unstable: standard errors become inflated, and the model struggles to attribute variance to a specific predictor, often arbitrarily splitting importance between redundant variables. This makes it difficult to determine whether the market is pricing the average water level, the extreme water level, or the frequency of inundation.

To resolve this, we utilized Lasso regression, which introduces an L1 regularization penalty to the objective function. We constructed a processing pipeline that first normalizes all inputs via the scikit-learn `StandardScaler` a critical prerequisite to ensure the penalty applies uniformly across features with differing magnitudes. The objective function minimizes the residual sum of squares subject to a constraint on the absolute size of the coefficients.

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^N (FMV_i - \sum_x \beta_j X_{ij})^2 + \lambda \sum_{j=1}^p |\beta_j|$$

Lasso forces the coefficients of redundant or non-informative features exactly to zero, giving us a standardized method for selecting an informative subset from our universe of features, which we display in Figure 4. By regressing these isolated risk signals against the price residuals, we aim to determine if we can reject the null hypothesis that flood risk is uncorrelated with the remaining price variance. If the market were correctly pricing this exposure, we would expect to reject the null, finding a strong negative relationship between risk and price residuals. Conversely, a failure to reject this null would imply that despite the physical reality of frequent inundation, the market has not yet integrated this specific chronic risk signal into property valuations. This residual analysis forms the basis of the results presented in Section 5.

## 5 Modeling Systemic Overpricing

### 5.1 Hedonic Pricing Model

To quantify the implicit value of environmental and structural attributes, we adopted a hedonic pricing framework which treats a property as a bundle of distinct characteristics (such as location, size, and risk profile) each contributing a marginal utility to the final transaction price. A defining characteristic of real estate financial data is its high variance and heavy-tailed distribution, where a small number of luxury properties can disproportionately skew statistical estimates (see Figure 5). Earlier studies have demonstrated that using a logarithmic model instead of a linear specification significantly improves

Signal Name	Notation / Formula	Description
<b>Freeboard Metrics (Vertical Safety Margin)</b>		
Freeboard (Min)	$F_{\min} = E - w_{\max}$	The vertical distance between property elevation ( $E$ ) and the annual maximum water level. Negative values imply inundation.
Freeboard (Mean)	$F_{\text{mean}} = E - \mu_w$	The distance between property elevation and the annual mean water level.
Freeboard (Median)	$F_{\text{med}} = E - w_{p50}$	Distance from property elevation to the median water level.
Freeboard (P90)	$F_{p90} = E - w_{p90}$	Distance from property elevation to the 90th percentile water level.
Freeboard (P99)	$F_{p99} = E - w_{p99}$	Distance from property elevation to the 99th percentile water level.
<b>Relative Risk Metrics</b>		
Risk Ratio (Max)	$R_{\max} = \frac{w_{\max}}{\max(E, 0.1)}$	Ratio of maximum water level to property elevation. Values $> 1$ indicate flooding.
Risk Ratio (Mean)	$R_{\text{mean}} = \frac{\mu_w}{\max(E, 0.1)}$	Ratio of mean water level to property elevation.
Inverse Distance Risk	$I_{\text{dist}} = \frac{1}{\max(F_{\min}, 0.1)}$	Inverse of the minimum freeboard; emphasizes properties near the breach threshold.
Z-Score Risk	$Z = \frac{E - \mu_w}{\sigma_w}$	Property elevation expressed as the number of standard deviations above the mean water level.
<b>Exceedance Probabilities &amp; Durations (Hourly)</b>		
Hours Above $X$ ft	$\sum \mathbb{I}(w_t > E - X)$	Total count of hours where water levels exceeded a safety buffer of $X$ feet below the property (for $X \in \{1, 2, 3, 5, 7, 10\}$ ).
Prob. Exceed 1 ft	$P(w > E - 1)$	The probability (fraction of the year) that water levels are within 1 ft of (or higher than) the property elevation.
Prob. Exceed 3 ft	$P(w > E - 3)$	The probability that water levels are within 3 ft of (or higher than) the property elevation.
<b>Exceedance Durations (Daily)</b>		
Days Above $X$ ft	$\sum \mathbb{I}(\max(w_d) > E - X)$	Total number of days where the daily maximum water level exceeded the safety buffer $X$ (for $X \in \{1, 2, 3\}$ ).
<b>Annual Water Statistics</b>		
Yearly Water Max	$w_{\max}$	The absolute maximum water level recorded in the year.
Yearly Water Mean	$\mu_w$	The arithmetic mean of all water level observations in the year.
Yearly Water Std	$\sigma_w$	The standard deviation of water levels for the year.

Table 1: Description of constructed flood risk signals.  $E$  represents the parcel elevation (NAVD88),  $w$  represents the water level time series, and  $\mathbb{I}$  is the indicator function.

the stability of coefficient estimates, particularly given that property prices tend to vary widely across locations and years (Avanijaa et al., 2021). By defining our target variable as  $\ln(Y_i)$ , we achieved a distribution that approximates normality after outlier processing. This transformation mitigates the influence of extreme values and allows for the interpretation of coefficients as percentage changes rather than fixed dollar amounts.

While traditional Ordinary Least Squares (OLS) regression provides straightforward interpretability, it often struggles with the complex, non-linear interactions inherent in housing markets. To address these limitations, we selected Extreme Gradient Boosting (XGBoost) as our primary modeling architecture. Recent housing market studies consistently find that tree-based models outperform linear regressions when relationships between features and prices are non-linear (Sharma et al., 2024). XGBoost is particularly effective on structured real estate datasets because it automatically captures

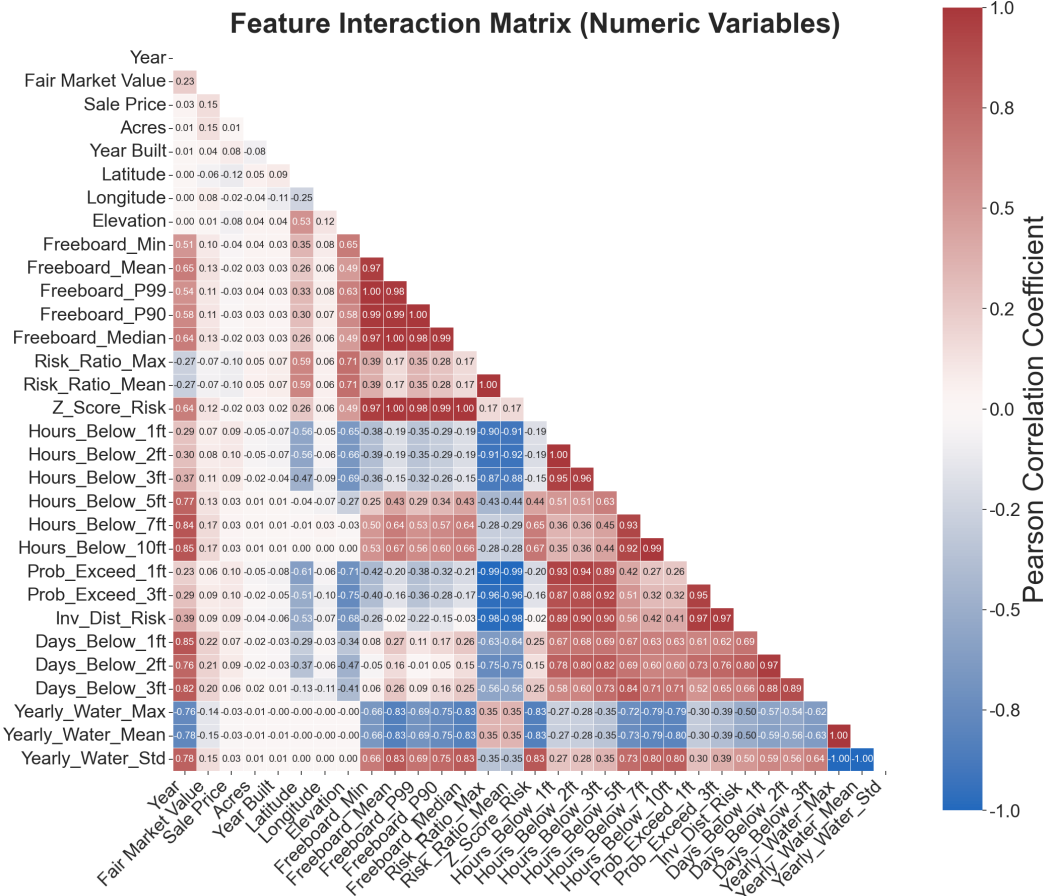


Figure 3: Correlation Matrix for engineered risk-signals

the interaction effects between location, structural features, and environmental variables without requiring manual specification. Furthermore, prior research indicates that XGBoost minimizes outlier influence compared to traditional regression models, a critical advantage in property markets that frequently deal with irregular transactions (Sharma et al., 2024).

We can simply formulate our pricing model as follows.

$$\ln(\text{FMV}_i) = S(\mathbf{X}_i, \mathbf{R}_i) + \epsilon_i$$

$S$  is the non-linear function estimated by the gradient boosting ensemble. This formulation supports a two-stage residual analysis framework. We first estimate a "baseline" valuation using only  $\mathbf{X}_i$  to isolate the price variance unrelated to standard property characteristics. In the second stage, we regress the excluded flood risk signals ( $\mathbf{R}_i$ ) directly against the resulting residuals ( $\epsilon_i$ ). This procedure allows us to explicitly test for information contribution: if the housing market is efficiently capitalizing flood risk, the risk signals ( $\mathbf{R}_i$ ) should demonstrate a statistically significant correlation with the price residuals

## 5.2 Training and Statistical Inference

To isolate the specific impact of flood risk on property valuation, we established a rigorous machine learning pipeline designed to control for standard non-linear hedonic price drivers. The training procedure focuses on estimating a Baseline valuation model using only standard physical and spatial attributes. This baseline model allows us to generate a set of price residuals for subsequent efficiency testing. Prior to model ingestion, the raw dataset underwent a series of transformations to stabilize

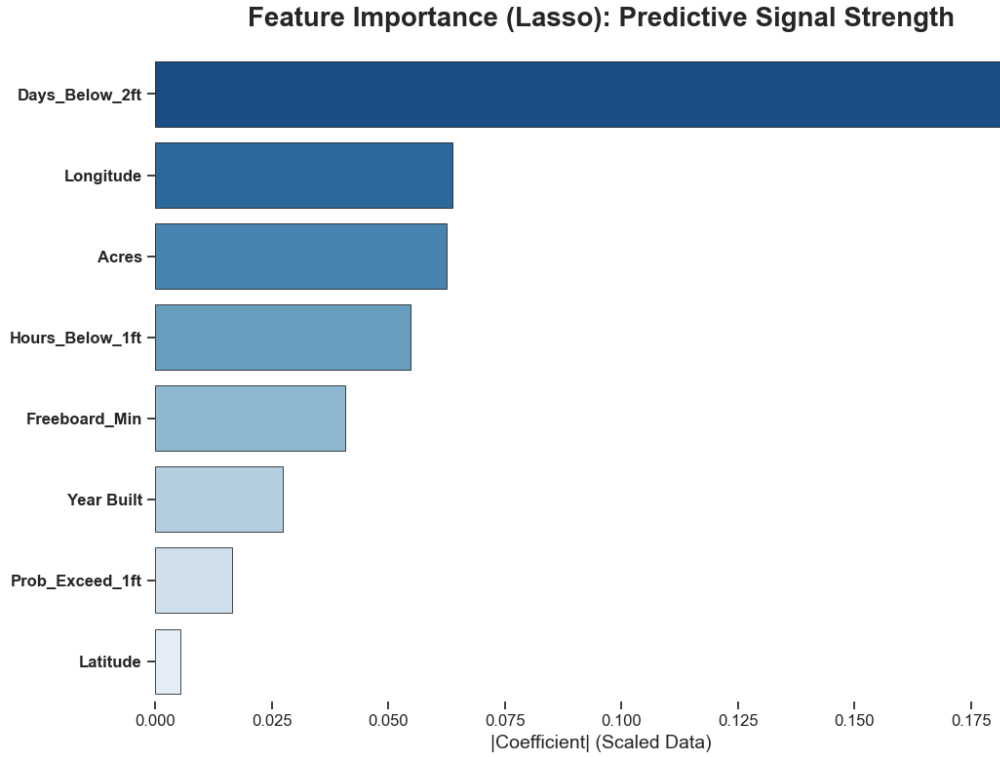


Figure 4: Correlation Matrix for engineered risk-signals

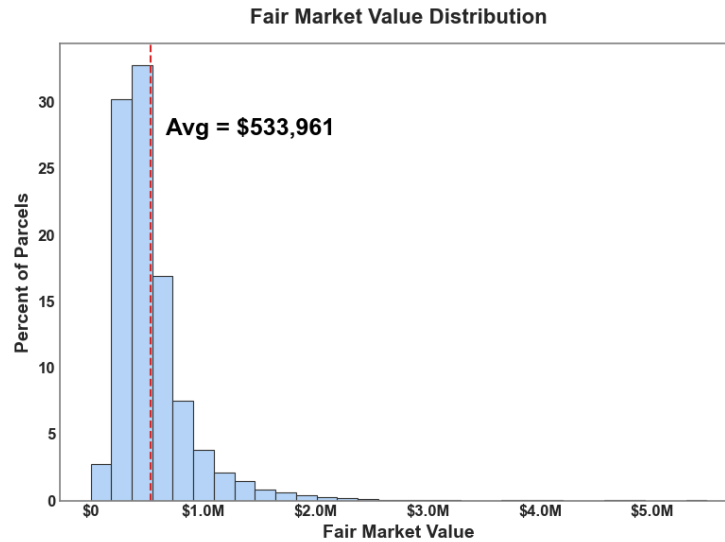


Figure 5: Distribution of Fair Market Value displays prominent Log-Normal skew

variance and improve convergence. Given the characteristic right-skewed distribution of real estate financial data and land measurements, we applied logarithmic transformations to both the dependent variable and the lot size feature. Additionally, we derived a property Age feature calculated as the difference between the transaction year and the year the structure was built. The data was strictly filtered to exclude non-arms-length transactions, such as zero or negative values. We then split the data into training and testing sets using an 80-20 split and a fixed random seed to ensure reproducibility.

To prevent feature dominance due to magnitude differences, such as Year versus Latitude, all input features were standardized to zero mean and unit variance using StandardScaler.

We employed Extreme Gradient Boosting, or XGBoost, as our primary regression estimator. Unlike traditional linear hedonic models like OLS, XGBoost naturally handles the complex, non-linear interactions between spatial variables and property attributes without requiring manual interaction term engineering. To ensure the model generalized well to unseen data, we implemented a 5-fold Cross-Validated Randomized Search over a robust hyperparameter space. The search spanned 50 iterations and optimized for the coefficient of determination across a specific grid. This grid included tree structure parameters such as max depth and minimum child weight, ensemble dynamics like estimators and learning rate, stochastic sampling rates, and regularization terms to penalize complexity.

A critical component of our methodology is the distinction between the Naive Baseline Model and the full feature space. To test for market efficiency, we first trained the model using only standard hedonic controls including log acres, age, latitude, longitude, and year. The residuals generated by this baseline model on the hold-out test set represent the variation in property prices not explained by size, age, or general location. In a perfectly efficient market that prices flood risk, these residuals should be strongly correlated with our omitted flood risk signals, such as Days Below 2ft. Conversely, if the residuals show no statistical relationship with the risk signals, it suggests the market has failed to capitalize this specific exposure into property values.

## 6 Analysis of Residuals

### 6.1 Regressing Flood-Signals against Residuals

To empirically test the efficiency of the Tybee Island housing market, we analyzed the relationship between our engineered flood risk signals and the price residuals derived from the baseline hedonic model. The theoretical framework for this test is straightforward. If the market efficiently capitalizes flood risk, then properties with higher exposure should trade at a discount relative to their structurally equivalent but safer peers. Consequently, the baseline model, which accounts for structural and locational attributes but omits flood data, should systematically overpredict the value of high-risk homes. This would result in a statistically significant negative correlation between the risk signals and the model residuals. We performed a series of Ordinary Least Squares regressions using the naive model residuals as the dependent variable and our key flood risk metrics as independent predictors. The results provide strong evidence against the efficient market hypothesis in this context.

We first examined `Days_Below_2ft`, which the Lasso analysis identified as the most informative feature in the dataset. As illustrated in Figure 6, the regression yields a coefficient of determination ( $R^2$ ) of just 0.0012. The p-value of 0.1429 exceeds standard significance thresholds, meaning we fail to reject the null hypothesis. There is no statistically detectable relationship between the frequency of tidal inundation and the unexplained variance in home prices. The regression line is effectively flat, indicating that errors in the baseline valuation model are randomly distributed rather than systematic functions of flood exposure.

This pattern persists across other definitions of risk. The regression of residuals against `Prob_Exceed_3ft`, shown in Figure 7, produced an even lower  $R^2$  of 0.0007 and a non-significant p-value of 0.2582. While the metric `Hours_Below_5ft` in Figure 8 did return a statistically significant p-value of 0.0023, the effect size is negligible with an  $R^2$  of only 0.0052. This suggests that while a trace mathematical relationship may exist, it lacks economic meaningfulness and explains less than 1% of the pricing error. The absence of a negative correlation leads to a clear conclusion regarding the current state of the Tybee Island real estate market. If buyers were penalizing properties for chronic flood exposure, the baseline model would consistently exhibit negative residuals for high-risk parcels. The fact that it does not implies that flood risk is currently orthogonal to price formation. The market appears to be pricing these assets based almost exclusively on their physical characteristics and general location, ignoring the granular, property-specific hydrological risks that our signals successfully capture. This disconnection confirms the presence of a valuation bubble driven by information asymmetry or risk insensitivity.

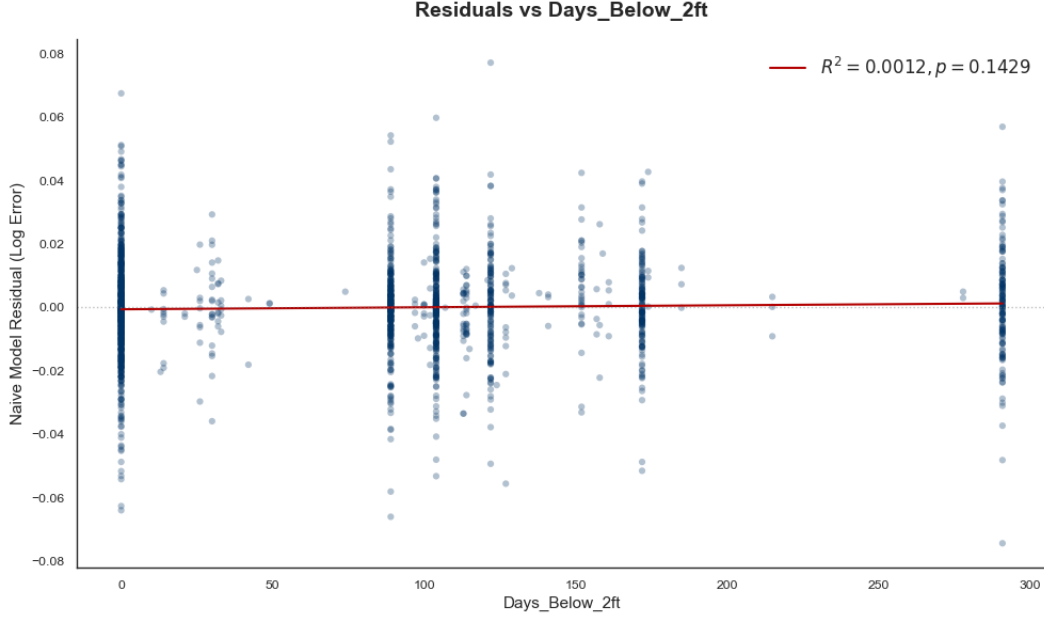


Figure 6: OLS Regression of Residuals v.s Days Below 2ft Signal

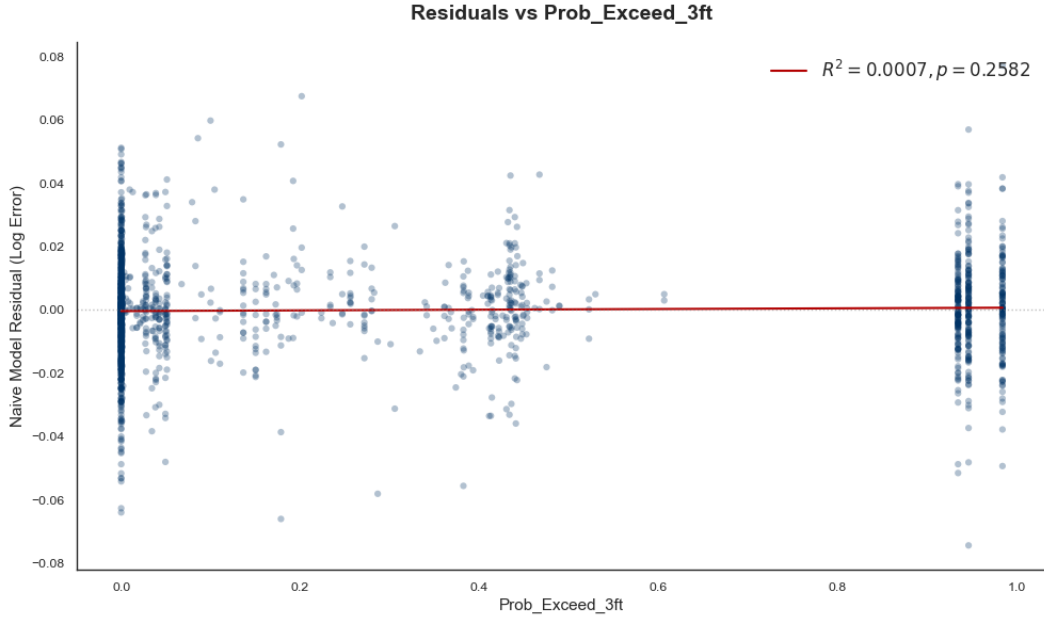


Figure 7: OLS Regression of Residuals v.s Prob. Exceed 3ft Signal

## 6.2 Over-valuation Thresholds

Following the regression analysis, we examined the distributional properties of the model residuals to identify specific instances of statistical overvaluation. While the aggregate market shows no systematic discount for flood risk, individual transactions exhibit significant variance. To quantify this, we standardized the logarithmic pricing errors into Z-scores, creating a normalized metric of deviation from fundamental value. We defined the bounds of "rational" pricing variation using a two-standard-deviation threshold ( $\mu \pm 2\sigma$ ). Under the assumption of a normally distributed error term, approximately 95% of transactions should fall within this confidence interval. Properties falling

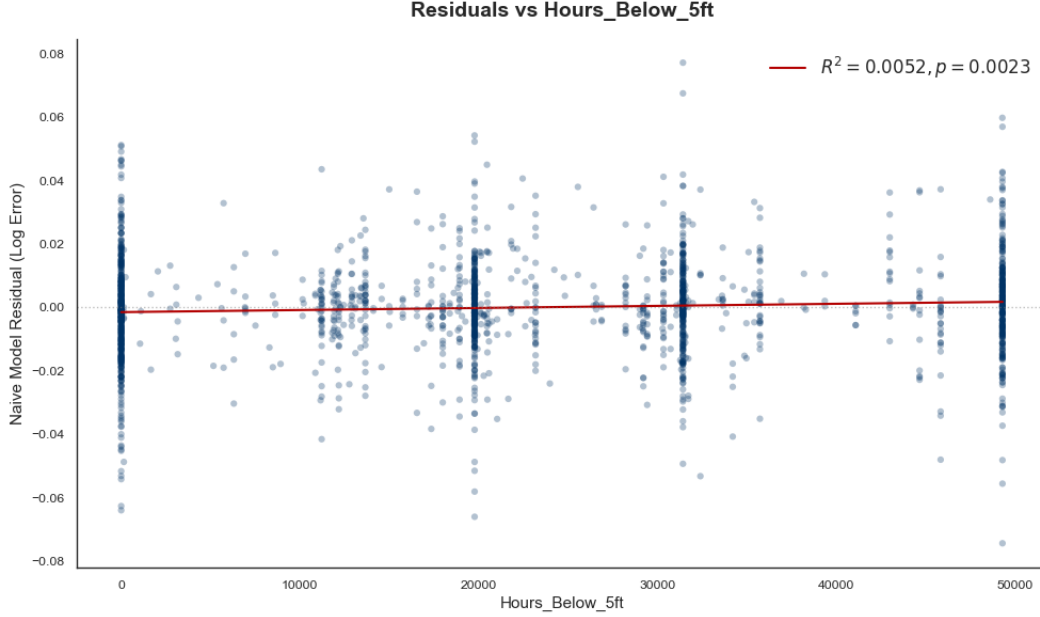


Figure 8: OLS Regression of Residuals v.s Hours Below 5ft Signal

Flood Risk Signal	Coefficient ( $R^2$ )	P-Value	Significance
Days Below 2ft	0.0012	0.1429	Not Significant
Hours Below 5ft	0.0052	0.0023	Significant ( $p < 0.01$ )
Hours Below 3ft	0.0015	0.1021	Not Significant
Prob Exceed 3ft	0.0007	0.2582	Not Significant
Freeboard (Min)	0.0006	0.3071	Not Significant
Freeboard (Mean)	0.0013	0.1268	Not Significant
Freeboard (Median)	0.0013	0.1304	Not Significant
Freeboard (P90)	0.0009	0.2028	Not Significant
Freeboard (P99)	0.0007	0.2543	Not Significant
Elevation	0.0007	0.2738	Not Significant

Table 2: Results of OLS regression where the dependent variable is the residual ( $\epsilon$ ) from the baseline hedonic model. Low  $R^2$  values indicate that flood risk signals explain virtually none of the pricing error, suggesting the market is not capitalizing this risk.

outside these bounds represent statistical anomalies. Overvalued properties are where ( $Z > +2$ ); transactions where the actual sale price significantly exceeded the model-implied valuation. These represent cases where the buyer paid a premium unsupported by the property's physical or spatial attributes.

### 6.3 Quantile Regression

To distinguish between normal market variance and significant pricing anomalies, we moved beyond point estimates (conditional means) to estimate the conditional distribution of housing prices. We defined a Rational Price Corridor: a 90% prediction interval derived exclusively from standard hedonic attributes (the Naive feature set). We employed Gradient Boosting Regressors optimized for quantile loss rather than squared error. Unlike standard least squares regression which estimates the conditional mean  $E[y|X]$ , quantile regression estimates the conditional median or other percentiles. To construct the bounds of our corridor, we trained two distinct models: one to predict the 95th percentile ( $Q_{0.95}$ ) and one to predict the 5th percentile ( $Q_{0.05}$ ) of the log-price distribution. The objective function minimized by the Gradient Boosting Regressor is the "pinball loss" (or tilted

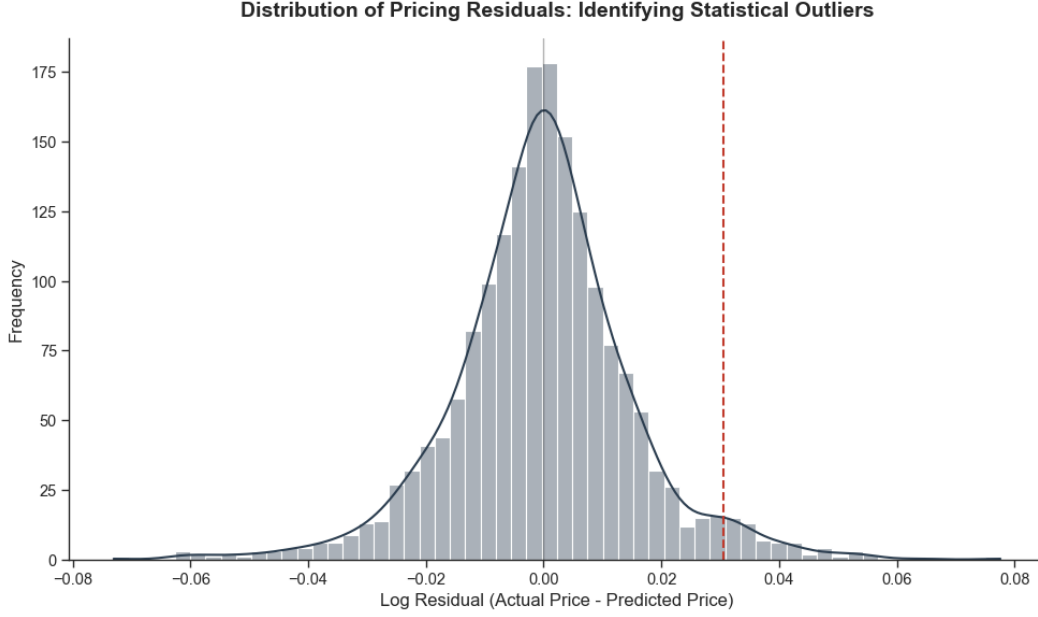


Figure 9: Statistically significant threshold for property overvaluation

absolute value loss). For a given quantile  $\tau \in (0, 1)$ , the loss function  $L_\tau$  for a prediction  $\hat{y}$  and actual value  $y$  is defined as follows.

$$\hat{Q}_\tau(x) = \operatorname{argmin}_{f \in \mathcal{F}} \sum_{i=1}^n (\tau - \mathbb{I}_{(FMV_i < f(x_i))}) (FMV_i - f(x_i))$$

We specified  $\tau = 0.95$  for the upper bound model ( $\hat{FMV}_{upper}$ ) and  $\tau = 0.05$  for the lower bound model ( $\hat{FMV}_{lower}$ ). The resulting Rational Price Corridor is the closed interval  $[\hat{FMV}_{lower}, \hat{FMV}_{upper}]$ . Properties falling within this interval represent transactions where the price can be explained by standard structural attributes (e.g., square footage, year built) within a reasonable margin of market volatility. However, properties falling outside these bounds signal significant mispricing. The results of this classification are visualized in Figure ?? . The gray band represents the 90% confidence interval. Points plotted in red indicate transactions exceeding the 95th percentile (Overpriced), while green points indicate transactions below the 5th percentile (Underpriced).

## 7 Conclusions

This study provides empirical evidence of a systemic market failure within the Tybee Island real estate market, confirming the existence of a localized "climate bubble." By integrating high-frequency hydrological telemetry from SECOORA sensors with granular parcel elevation data, we successfully engineered a suite of hydrostatic risk signals that capture the chronic reality of tidal inundation far more precisely than static federal flood maps. Despite the physical severity of these risks, our residual analysis demonstrates that they are statistically invisible in current transaction prices.

Ultimately, while our analysis is rooted in the specific hydrology of Tybee Island, the methodology presented here is designed for broad generalization across the Atlantic and Gulf Coasts. We demonstrate that the integration of open-source hydrological telemetry with machine learning is not merely an academic exercise, but a scalable necessity. By providing a robust proof-of-concept for real-time risk inference, our infrastructure offers municipalities, insurers, and homeowners a critical tool to navigate an increasingly opaque market. As institutional transparency regarding climate hazards diminishes, such data-driven frameworks will be essential in restoring market efficiency and ensuring that the financial valuation of coastal real estate aligns with its physical future.

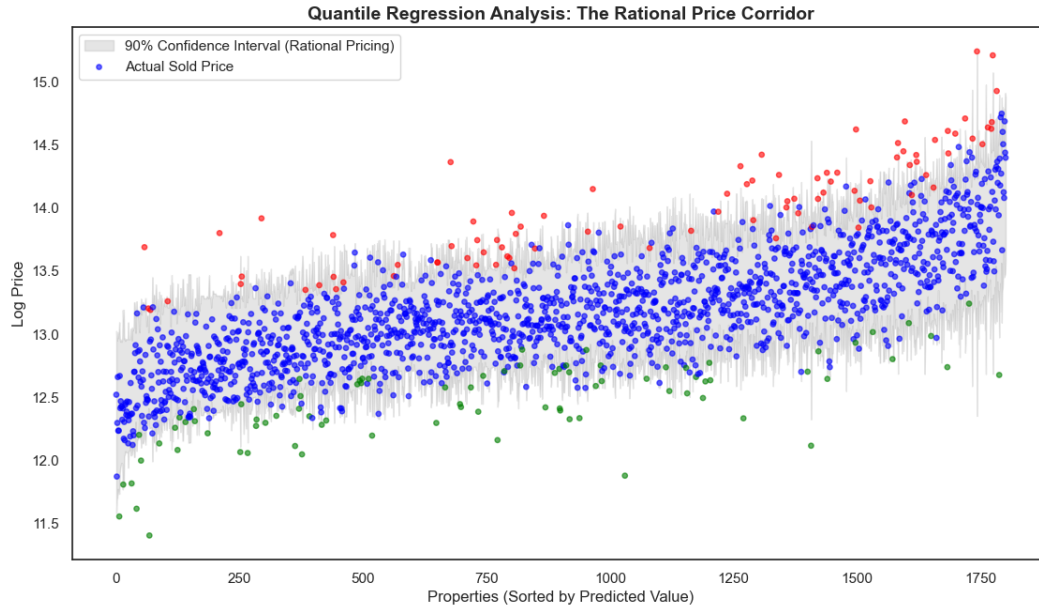


Figure 10: Rational Price Corridor

## References

- Avanijaa, J. (2021). Prediction of house price using machine learning algorithms. *International Journal of Engineering Research & Technology (IJERT)*, 10(6).
- Gourevitch, J. D., Kousky, C., Liao, Y., Nolte, C., Pollack, A. B., Porter, J. R., & Weill, J. A. (2023). Unpriced climate risk and the potential consequences of overvaluation in US housing markets. *Nature Climate Change*, 13(3), 250–257. <https://doi.org/10.1038/s41558-023-01594-8>
- The Guardian. (2025, December). Zillow removes climate risk scores from listings following industry pressure. *The Guardian*. <https://www.theguardian.com>
- Hino, M., & Burke, M. (2021). The effect of information about climate risk on property values. *Proceedings of the National Academy of Sciences*, 118(17), e2003374118. <https://doi.org/10.1073/pnas.2003374118>
- Milliman. (2022). Milliman flood insurance market analysis. Milliman, Inc.
- Realtor.com. (2025). 2025 Housing market climate risk report. News Corp.
- SECOORA. (2025). Water level data from sensor gt-envsense-069 [Data set]. Southeast Coastal Ocean Observing Regional Association. <https://api.sealevelsensors.org/v1.0>
- Sharma, A., & Associates. (2024). Comparative analysis of tree-based ensemble models versus linear regression for real estate valuation. *Journal of Real Estate Research*.