**REVISED COMPREHENSIVE PROPOSAL**

***Harnessing Ocean and Weather Data to Predict Coastal San Diego Ozone: A Tiered Machine Learning Approach for Community Health***

**1. INTRODUCTION**

Ground-level ozone pollution poses significant health risks to San Diego's coastal communities, particularly affecting children, elderly residents, and those with respiratory conditions. Current air quality forecasts rely primarily on land-based meteorological data, potentially missing crucial ocean-atmosphere interactions that influence coastal ozone formation and transport. This project proposes a novel, tiered machine learning approach that systematically integrates marine upwelling indices and land-ocean temperature contrasts with traditional meteorological predictors to improve ozone forecasting accuracy and community health protection.

**2. REVISED HYPOTHESIS**

**"Incorporating marine upwelling indices (CUTI/BEUTI) and land-ocean temperature contrasts into machine learning models will significantly improve the accuracy of predicting daily maximum ground-level ozone concentrations at coastal San Diego air monitoring sites compared to using land-based meteorological variables alone."**

**Specific Sub-Hypotheses:**

1. **CORE HYPOTHESIS:** Adding upwelling variables (CUTI, BEUTI) to standard meteorological predictors will increase ozone prediction accuracy by 10-20% (R² improvement of 0.05-0.15).
2. **MECHANISTIC HYPOTHESIS:** Strong upwelling events and large land-sea temperature contrasts will create distinct 'ozone regimes'—with lower ozone concentrations when marine influence is strong (high CUTI, large temperature gradients) and higher ozone when marine influence is weak (low CUTI, stagnant conditions).
3. **TEMPORAL HYPOTHESIS:** Ocean effects on ozone will show lag relationships, with upwelling events 1-3 days prior influencing current ozone levels through persistent marine layer effects.

**3. EXPERIMENTAL DESIGN WITH TIERED FEATURE APPROACH**

**Variables:**

* **Independent (Predictors):** Tiered feature set based on scientific priority (69 total variables documented)
* **Dependent:** Daily maximum 8-hour ozone concentration (ppb)

**Experimental Groups:**

* **CONTROL MODEL:** Land-based meteorology + temporal features only
* **EXPERIMENTAL MODEL:** Land meteorology + ocean upwelling + land-ocean interactions

**Three-Tier Feature Selection Strategy:**

**TIER 1 - ESSENTIAL FEATURES (Minimal Model, n=15-20):**

* **Target:** ozone\_ppb
* **Core Predictors:** ozone\_lag\_1, tmax, tavg, CUTI, month\_sin, month\_cos, wspd
* **Temporal:** year, month, dayofweek
* **Land-Ocean:** Tmax\_inland, land\_sea\_temp\_diff
* **Additional:** tmin, pres, ozone\_roll\_mean\_7

**TIER 2 - ENHANCED MODEL (Standard Model, n=25-35):**

* **All Tier 1 features plus:**
* **Ocean Lags:** CUTI\_lag1, CUTI\_lag3, CUTI\_roll7\_mean
* **Marine Effects:** thermal\_stability, marine\_layer\_presence
* **Upwelling:** BEUTI, strong\_upwelling\_event
* **Meteorology:** tsun, temp\_range

**TIER 3 - COMPREHENSIVE MODEL (Advanced Model, n=40-50):**

* **All Tier 1&2 features plus:**
* **Extended Lags:** CUTI\_lag7, ozone\_roll\_std\_7, sst\_value\_sst\_lag1
* **Additional Ocean:** BEUTI\_lag1, Bakun indices, sst\_anomaly
* **Events:** ozone\_exceed\_70, heatwave\_flag
* **Spatial:** distance\_to\_coast\_km, region\_type
* **Interaction Terms:** tmax×CUTI, wspd×land\_sea\_temp\_diff

**4. METHODOLOGY - PROGRESSIVE MODEL BUILDING**

**Data Preparation:**

* Clean and validate all 69 documented variables
* Handle missing values using domain-appropriate methods
* Create engineered features (lags, rolling statistics, interactions)
* Split data: 2020-2022 (training), 2023 (testing)

**Model Development (Progressive Complexity):**

**STAGE 1 - Baseline Models:**

* Control Model: Tier 1 features minus ocean variables (n≈12)
* Minimal Experimental: Tier 1 complete features (n≈18)
* Algorithm: Random Forest, XGBoost
* Evaluation: R², RMSE, MAE on test set

**STAGE 2 - Enhanced Models:**

* Standard Experimental: Tier 1 + Tier 2 features (n≈30)
* Feature Selection: Recursive Feature Elimination
* Cross-validation: 5-fold temporal cross-validation
* Hyperparameter tuning: Grid search

**STAGE 3 - Advanced Models (if time/resources permit):**

* Comprehensive Model: Tier 1 + 2 + selected Tier 3 (n≈45)
* Dimensionality Reduction: PCA on correlated upwelling variables
* Advanced algorithms: Neural networks, ensemble methods

**Regime Analysis:**

* Apply k-means clustering (k=3-4) to all daily observations
* Use PCA for visualization of ozone regimes
* Characterize clusters by upwelling strength, marine layer presence
* Validate regime interpretation with meteorological expertise

**Performance Evaluation:**

* Primary Metric: R² improvement (experimental vs control)
* Secondary Metrics: RMSE, MAE, especially for high ozone days (>70 ppb)
* Feature Importance: SHAP values, permutation importance
* Statistical Significance: Paired t-tests between model performances

**5. EXPECTED RESULTS BY MODEL TIER**

**TIER 1 RESULTS (Minimal Model):**

* Control Model R²: 0.60-0.70 (typical for meteorology-only)
* Experimental Model R²: 0.65-0.75 (+5-10% improvement)
* Key Finding: CUTI and land\_sea\_temp\_diff among top 5 predictors
* Feature Importance: ozone\_lag\_1 > tmax > CUTI > month\_sin > tavg

**TIER 2 RESULTS (Enhanced Model):**

* Control Model R²: 0.65-0.72 (with more meteorological features)
* Experimental Model R²: 0.72-0.80 (+7-15% improvement)
* Key Finding: CUTI lags (1-3 days) improve prediction of marine layer events
* Marine layer proxies significantly improve low-ozone day predictions

**TIER 3 RESULTS (Comprehensive Model):**

* Control Model R²: 0.70-0.75 (diminishing returns from complexity)
* Experimental Model R²: 0.75-0.82 (+5-10% additional improvement)
* Key Finding: Interaction terms (tmax×CUTI) capture non-linear effects
* Event flags improve extreme ozone prediction

**REGIME ANALYSIS RESULTS:**

* Cluster 1: "Marine-Dominated Days" (high CUTI, low ozone, strong gradients)
* Cluster 2: "Stagnant/Hot Days" (low CUTI, high ozone, weak marine influence)
* Cluster 3: "Transitional Days" (moderate conditions, variable ozone)
* Validation: Clusters align with known meteorological patterns

**6. SCIENTIFIC AND SOCIETAL SIGNIFICANCE**

**Scientific Contributions:**

1. First high school study to quantify ocean upwelling effects on coastal ozone
2. Novel application of CUTI/BEUTI indices to air quality prediction
3. Systematic tiered approach to feature selection in environmental ML
4. Validation of land-sea temperature gradient importance in ozone formation
5. Discovery of predictive "ozone regimes" linked to marine conditions

**Community Health Impact:**

* Improved next-day ozone forecasts for vulnerable populations
* Better understanding of when ocean conditions protect air quality
* Actionable insights for schools/parents on outdoor activity planning
* Evidence-based support for marine protection as air quality co-benefit

**Practical Applications:**

* Enhanced early warning systems for high ozone days
* Improved seasonal ozone forecasting using ocean condition forecasts
* Better targeted public health messaging based on marine layer presence
* Support for coastal zone management considering air quality benefits

**7. ACKNOWLEDGED LIMITATIONS AND SCIENTIFIC HONESTY**

**Data Limitations:**

* Sea surface temperature gaps near coast due to satellite limitations
* Upwelling indices represent regional (33°N) not local conditions
* Weather data from limited inland stations for temperature contrasts
* Ozone data availability varies by site and time period

**Methodological Limitations:**

* Correlation does not imply causation in observational study
* Model performance limited by data quality and completeness
* Feature selection may miss important non-linear interactions
* Regime analysis subjective in optimal cluster number selection

**Honest Assessment:**

* Ocean effects may be smaller than expected if already captured by meteorology
* Some "significant" results may be due to multiple testing without correction
* Model improvements might not be practically meaningful despite statistical significance
* Feature importance may vary significantly across different time periods or sites