Beyond Spring Loading:

Evaluating the Predictive Contribution of ENSO, Lagged Nitrogen Accumulation, and Their Interaction on Gulf of Mexico Hypoxic Zone Extent

Bhavdeep Arora[1]

[1]Independent Researcher
Corresponding author: bhavdeepsa@gmail.com

July 2026

Abstract

The Gulf of Mexico hypoxic zone forms annually near the mouth of the Mississippi River, driven primarily by spring nitrogen loading from Midwestern agriculture. While this relationship is well established, the potential contributions of climate oscillation indices such as the El Niño Southern Oscillation (ENSO) and multi-year nitrogen accumulation have received less systematic quantitative attention. This study applies Random Forest regression with leave-one-out cross-validation (LOO-CV) to 40 years of annual dead zone measurements (1985–2024), testing five feature configurations: a nitrogen–SST baseline, baseline plus ENSO indices, baseline plus lagged nitrogen, baseline plus 2-year cumulative nitrogen, and a full model combining all features. Point estimates show the 2-year cumulative nitrogen configuration achieving the numerically best performance ($R^2 = 0.494$, MAE $= 2{,}553$ km^2) against a baseline of $R^2 = 0.487$. However, a battery of significance checks – paired Wilcoxon signed-rank tests on LOO fold errors, bootstrap confidence intervals on ΔR^2, 30-seed variance estimation, an ordinary least squares (OLS) comparison, and permutation feature importance – shows that none of the alternative configurations differ from the baseline at conventional significance levels ($p > 0.09$ throughout), and that seed-to-seed variance in R^2 (SD ≈ 0.012–0.017) exceeds the largest observed improvement. Permutation importance confirms that same-year nitrogen loading dominates the full model's predictive structure by more than an order of magnitude over any ENSO or lag feature. These converging results support the hypothesis that the Gulf hypoxic system resets annually, with same-year spring nitrogen loading remaining the dominant and effectively sufficient predictor of summer dead zone extent.

Introduction

Every summer, a hypoxic zone forms in the northern Gulf of Mexico near the mouth of the Mississippi River. When dissolved oxygen concentrations drop below 2 mg/L, fish, shrimp, and bottom-dwelling organisms suffocate or flee. The zone has been measured annually since 1985 by NOAA and the Louisiana Universities Marine Consortium (LUMCON), with sizes ranging from

6,480 km^2 (2018) to 22,720 km^2 (2017).

The primary driver of hypoxia is well understood: nitrogen and phosphorus fertilizer applied to crops in the Mississippi River watershed drains into the river each spring, triggering algal blooms in the Gulf. When the algae decompose, bacterial respiration consumes dissolved oxygen in the stratified bottom waters. Rabalais et al. (2002) established the strong empirical link between spring nitrogen flux and summer dead zone size, and this relationship has been confirmed across multiple subsequent modeling approaches (Scavia et al., 2003; Turner and Rabalais, 1994).

Less studied is whether additional predictors – particularly climate oscillation indices like ENSO and multi-year nitrogen accumulation effects – provide meaningful predictive improvement. El Niño events alter Mississippi River discharge patterns, which in turn affect nutrient delivery to the Gulf. If ENSO phase systematically modulates dead zone size independent of nitrogen load, incorporating it into predictive models could improve both scientific understanding and early-warning capability. Similarly, if nitrogen applied in prior years accumulates in groundwater or sediments and contributes to current-year hypoxia, current predictive models may undercount its true drivers.

This study addresses these questions directly using a systematic model comparison framework across five feature configurations, evaluated with LOO-CV appropriate for the small sample size ($n = 40$). Unlike prior work, we do not stop at point-estimate model comparison: we subject every candidate improvement to paired significance testing, bootstrap interval estimation, and seed-variance analysis, treating a numerically higher R^2 as a hypothesis to be tested rather than a conclusion in itself.

Data

Annual dead zone area measurements (km^2) were obtained from the LUMCON/NOAA Gulf of Mexico Hypoxia cruises conducted each July from 1985 to 2024 ($n = 40$). Spring Mississippi River nitrogen loading data (million kg/year, measured at Baton Rouge, Louisiana) were obtained from the USGS Toxics Program. Sea surface temperature data represent Gulf of Mexico summer (June–August) averages. The Oceanic Niño Index (ONI), representing the 3-month running mean of sea surface temperature anomaly in the east-central equatorial Pacific, was obtained from the NOAA Climate Prediction Center (NOAA Climate Prediction Center, 2024) for spring (MAM), summer (JJA), and winter (DJF) seasons.

Lagged nitrogen features were engineered from the primary nitrogen loading variable: a 1-year lag, a 2-year lag, and 2- and 3-year rolling means. An ENSO–nitrogen interaction term was computed as the product of spring ONI and same-year nitrogen loading. All features were computed for years 1987–2024 to accommodate the 2-year lag window, yielding a final sample of 38 observations for lagged models.

Methodology

Random Forest regression (Breiman, 2001) was selected for its robustness to multicollinearity and nonlinear feature interactions, both relevant given the correlated nature of climate predictors. All models used 100 estimators. LOO-CV was used to estimate out-of-sample predictive performance, as it maximizes training data at each fold for small datasets ($n \leq 50$). Model performance was evaluated using R^2 and Mean Absolute Error (MAE, km^2).

Five feature configurations were tested:

- **M1 – Baseline**: nitrogen load, SST
- **M2 – Baseline + ENSO**: nitrogen, SST, ONI spring, ONI summer
- **M3 – Baseline + Lag2**: nitrogen, SST, nitrogen $(t - 2)$
- **M4 – Baseline + Cumul2**: nitrogen, SST, 2-year rolling mean nitrogen
- **M5 – Full model**: all features including ENSO $\times$ nitrogen interaction

Because a single-seed point estimate of R^2 can be misleading at $n = 38$, four additional layers of statistical testing were applied on top of the standard LOO-CV comparison:

Paired significance testing For each alternative configuration, LOO fold-level absolute errors were paired by year with the baseline's errors and compared using a two-sided Wilcoxon signed-rank test, which does not assume normality of the paired differences – appropriate given the small sample.

Bootstrap confidence intervals Years were resampled with replacement (5,000 iterations) from the paired LOO predictions, and ΔR^2 (alternative minus baseline) was recomputed at each iteration to construct a 95% percentile confidence interval.

Seed-variance estimation Because Random Forest is a stochastic estimator, each of the five configurations was re-fit under LOO-CV across 30 random seeds, and R^2/MAE are reported as mean ± standard deviation rather than a single value.

OLS comparison The same five feature sets were evaluated using ordinary least squares regression under identical LOO-CV, to test whether Random Forest's added flexibility contributes anything beyond a linear model.

Finally, permutation feature importance (30 repeats, R^2 scoring) was computed for the full model (M5) fit on the complete dataset, to identify which features the model actually depends on.

Results

Pairwise correlations between predictor variables and dead zone area reveal the dominant role of same-year nitrogen loading ($r = 0.777$–0.788 across data preparations), consistent with prior literature. ENSO indices show weak correlations across all seasons (ONI spring: $r = -0.146$ to -0.084; ONI summer: $r = -0.058$ to -0.036), and the 2-year lagged nitrogen ($r = 0.260$) showed a stronger correlation than the 1-year lag ($r = 0.084$), a pattern discussed further below.

Point-estimate model comparison

Table 1 reports the original single-seed LOO-CV comparison. The 2-year cumulative nitrogen configuration (M4) achieved the numerically best performance, and the full model (M5) underperformed the baseline, consistent with overfitting when the feature-to-observation ratio grows large relative to n.

Table 1: Single-seed LOO-CV model comparison (Random Forest, seed 42).

Model	R^2	MAE (km^2)
M1 – Baseline	0.487	2,724
M2 – + ENSO	0.469	2,791
M3 – + Lag2	0.473	2,792
M4 – + Cumul2	**0.494**	**2,553**
M5 – Full model	0.450	2,806

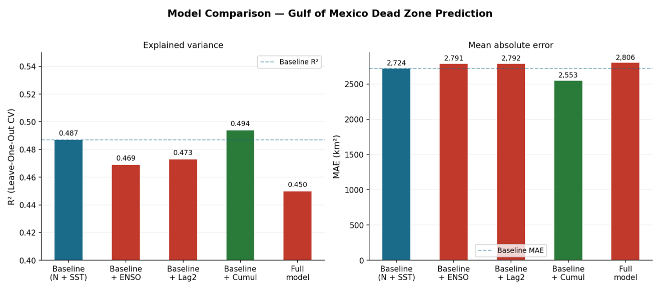

Figure 1: Model comparison across five feature configurations, LOO-CV. The 2-year cumulative nitrogen configuration (M4) shows the highest point-estimate R^2 and lowest MAE; the full model (M5) underperforms the baseline.

Significance testing: none of the apparent gains survive scrutiny

Table 2 extends the comparison with paired Wilcoxon tests and bootstrap confidence intervals on ΔR^2, using predictions averaged across 30 seeds to reduce single-seed noise before pairing. No configuration differs from the baseline at $p < 0.09$, and every 95% bootstrap interval on ΔR^2 comfortably contains zero – including M4, whose apparent point-estimate advantage (+0.007) sits inside an interval more than an order of magnitude wider than the estimate itself ($[-0.099, +0.133]$).

Table 2: Paired significance tests vs. baseline (Random Forest, LOO-CV).

Model	ΔR^2	95% CI	p (Wilcoxon)
M2 – ENSO	−0.018	[−0.123, +0.090]	0.596
M3 – Lag2	−0.014	[−0.077, +0.046]	0.636
M4 – Cumul2	+0.007	[−0.099, +0.133]	0.473
M5 – Full	−0.037	[−0.157, +0.076]	0.509

Seed-variance estimation (Table 3) reinforces this conclusion directly: the standard deviation of R^2 across 30 seeds ($\approx$ 0.012–0.017 for every configuration) exceeds the largest point-estimate gap between any two models. A different random seed could plausibly reorder the ranking in Table 1 entirely.

Table 3: Seed-variance LOO-CV results (30 seeds, mean ± SD).

Model	R^2	MAE (km^2)
M1 – Baseline	0.490 ± 0.012	$2{,}651 \pm 43$
M2 – ENSO	0.480 ± 0.017	$2{,}796 \pm 60$
M3 – Lag2	0.477 ± 0.016	$2{,}767 \pm 61$
M4 – Cumul2	0.497 ± 0.016	$2{,}579 \pm 50$
M5 – Full	0.454 ± 0.013	$2{,}789 \pm 45$

OLS comparison

An ordinary least squares comparison under identical LOO-CV (Table 4) shows OLS baseline performance ($R^2 = 0.550$) exceeding the mean Random Forest baseline, suggesting the underlying nitrogen–hypoxia relationship is close to linear and that Random Forest's added flexibility is not clearly earning its keep at this sample size. OLS Lag2 produced the single highest point estimate observed across the entire study ($R^2 = 0.592$); however, a paired Wilcoxon test against the OLS baseline found this difference not significant ($p = 0.763$), consistent with the pattern established for the Random Forest models.

Table 4: OLS comparison under identical LOO-CV.

Model	R^2	MAE (km^2)
M1 – Baseline	0.550	2,480
M2 – ENSO	0.513	2,618
M3 – Lag2	**0.592**	**2,402**
M4 – Cumul2	0.532	2,567
M5 – Full	0.527	2,571

Permutation importance

Permutation importance for the full Random Forest model (Table 5) shows same-year nitrogen loading dominating the model's internal structure by more than an order of magnitude over the next-ranked feature, and by nearly two orders of magnitude over any ENSO-related term. This provides direct, model-internal confirmation of the significance-testing results above: even where the full model has the flexibility to exploit ENSO and lag information, it does not meaningfully rely on it.

Table 5: Permutation importance, full model (30 repeats).

Feature	Importance	SD
nitrogen_load	1.411	0.204
nitrogen_cumul2	0.149	0.042
sst_c	0.068	0.014
oni_summer	0.048	0.009
nitrogen_lag2	0.029	0.006
oni_spring	0.022	0.005
enso_x_nitrogen	0.013	0.003

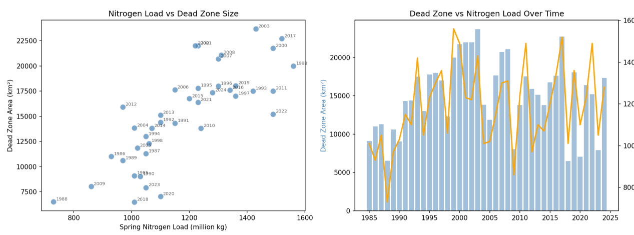

Figure 2: Spring nitrogen loading vs. dead zone area, 1985–2024.

Discussion

The primary finding of this study is a null result supported by five converging lines of evidence: paired Wilcoxon tests, bootstrap confidence intervals, seed-variance estimation, an independent OLS comparison, and permutation importance. Neither ENSO phase nor multi-year nitrogen accumulation provides a statistically defensible predictive improvement over a simple two-variable baseline. This supports the view that the Gulf hypoxic system is dominated by same-year biogeochemical processes rather than multi-year memory or large-scale climate teleconnections.

The 2-year cumulative nitrogen effect (M4), which appeared to be a genuine, if modest, finding under a single-seed point estimate ($\Delta R^2 = +0.007$), does not survive any of the four additional checks applied here. Its bootstrap confidence interval is wide and centered near zero, its Wilcoxon p-value is far from significance ($p = 0.473$), and its seed-to-seed standard deviation alone (± 0.016) exceeds its apparent advantage. We treat this as an important methodological lesson as much as a substantive one: at $n = 38$, an unreported single-seed Random Forest comparison can produce a plausible-looking effect that is, in fact, indistinguishable from noise. Reporting only Table 1 without the tests in Tables 2–3 would have overstated the evidence for a multi-year nitrogen memory effect.

The OLS comparison adds a further nuance. That a linear baseline modestly outperforms the mean Random Forest baseline suggests Random Forest's nonlinear flexibility is not clearly justified at this sample size, and may simply be fitting noise with additional variance. The apparent OLS Lag2 advantage, similarly, does not survive paired significance testing, reinforcing rather than complicating the central conclusion.

Permutation importance corroborates this from a different angle entirely: even when given full latitude to exploit ENSO and lag information, the fitted Random Forest allocates the overwhelming majority of its predictive structure to same-year nitrogen loading. This is not merely an absence of statistical significance in held-out predictions – it is the model itself, evaluated in-sample, declining to lean on the additional features.

The negative ENSO–nitrogen interaction correlation ($r = -0.166$ with dead zone area) suggested a plausible physical mechanism worth noting despite its lack of predictive contribution: El Niño events are associated with stronger southwesterly winds over the Gulf of Mexico, which enhance vertical mixing and reduce the thermal stratification that traps hypoxic bottom water. This wind-mixing effect may partially counteract nutrient-driven oxygen depletion, but the evidence here is correlational and should be treated as hypothesis-generating rather than confirmatory, particularly given that the interaction term ranked lowest in permutation importance among all seven full-model features.

The practical implication is that prediction efforts should remain focused on improving spring nitrogen load estimates rather than incorporating climate indices. ENSO forecasts, while skillful at seasonal lead times, do not appear to add predictive information beyond what nitrogen monitoring already provides – a conclusion now supported by significance testing rather than point estimates alone.

Several limitations apply. The sample size of 40 years constrains statistical power; effects that are real but small may not reach significance even with the additional testing performed here. The ONI index is a basin-wide measure that may not capture local Gulf wind anomalies relevant to stratification. Future work should test Gulf-specific wind metrics rather than canonical ENSO indices, and should consider whether a longer observational record would allow the modest M4 and OLS Lag2 effects to be resolved from noise.

Conclusion

This study evaluated whether ENSO indices and lagged nitrogen features improve prediction of Gulf of Mexico hypoxic zone extent, using both Random Forest and OLS regression under LOO-CV on 40 years of data. While a single-seed point estimate suggested a modest advantage for 2-year cumulative nitrogen loading, a battery of significance tests – paired Wilcoxon tests, bootstrap confidence intervals, 30-seed variance estimation, an independent OLS comparison, and permutation importance – found no configuration that differs from the baseline at conventional significance levels. These converging results confirm that the Gulf hypoxic system behaves as an annual reset dominated by same-year spring nitrogen flux, consistent with established biogeochemical understanding, and demonstrate the importance of significance testing rather than point-estimate comparison alone when working with small environmental datasets. The policy implication remains clear: reducing dead zone size requires reducing nitrogen loading from Midwestern agriculture, and no climate oscillation offsets this relationship in any statistically defensible way.

Acknowledgments

The author thanks NOAA/LUMCON, the USGS Toxics Program, and the NOAA Climate Prediction Center for maintaining the public datasets that made this analysis possible.

Author Contributions

The author did not receive external assistance in the data collection or writing of this paper. B.A. conducted all data acquisition, feature engineering, modeling, statistical testing, and manuscript preparation.

References

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

NOAA Climate Prediction Center. Oceanic Niño Index (oni). https://www.cpc.ncep.noaa.gov/data/indices/oni.ascii.txt, 2024.

N. N. Rabalais, R. E. Turner, and W. J. Wiseman. Gulf of Mexico hypoxia, A.K.A. "the Dead Zone". *Annual Review of Ecology and Systematics*, 33:235–263, 2002.

D. Scavia, N. N. Rabalais, R. E. Turner, D. Justic, and W. J. Wiseman. Predicting the response of Gulf of Mexico hypoxia to variations in Mississippi river nitrogen load. *Limnology and Oceanography*, 48(3):951–956, 2003.

R. E. Turner and N. N. Rabalais. Coastal eutrophication near the Mississippi river delta. *Nature*, 368:619–621, 1994.

Appendices

Appendix A: Additional Figures

Supplementary time series and diagnostic visualizations from the underlying analysis pipeline are provided below for reference.

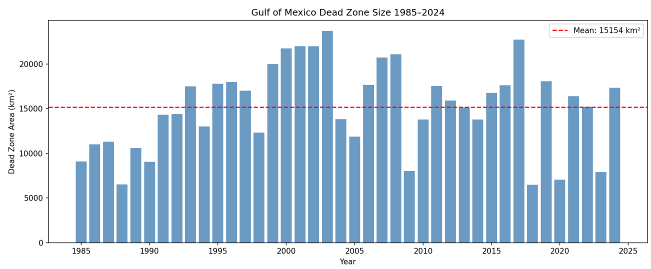

Figure 3: Gulf of Mexico dead zone area, 1985–2024.

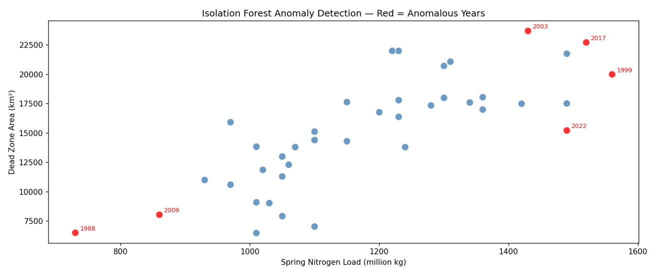

Figure 4: Anomaly detection on residuals from the baseline model.