# Predicting Bioluminescent Algal Blooms Using Hybrid ”EDM-LSTMs”

Athulith Paraselli
*Halıcıoğlu Data Science Institute*
*University of California San Diego*
La Jolla, United States of America
aparaselli@ucsd.edu

Ciro Zhang
*Halıcıoğlu Data Science Institute*
*University of California San Diego*
La Jolla, United States of America
ciz001@ucsd.edu

Esther Chung
*Halıcıoğlu Data Science Institute*
*University of California San Diego*
La Jolla, United States of America
esc005@ucsd.edu

Nian-Nian Wang
*Halıcıoğlu Data Science Institute*
*University of California San Diego*
La Jolla, United States of America
niw002@ucsd.edu

***Abstract*—Off the coast of Southern California, bioluminescent waves can be observed due to population blooms of “bioluminescent” dinoflagellates such as Lingulodinium polyedra. Forecasting these blooms has proven challenging due to the chaotic, non-linear nature of algal populations. Neural networks tend to have low prediction accuracies on chaotic systems due to their tendency to memorize patterns rather than learn generalizable dynamics. Furthermore, in algal bloom forecasting, many datasets contain relatively sparse and infrequent sampling, which further limits the neural network’s ability to capture the underlying dynamics. In this study, we utilize a Convolutional Neural Network–Long Short-Term Memory (CNN-LSTM) architecture, alongside Empirical Dynamic Modeling (EDM)-based embeddings, to enable the model to learn about the spatial and temporal dynamics of Lingulodinium polyedra. When predicting bioluminescent algal blooms, our hybrid EDM–LSTM model shows a $\tilde{1}2\%$ improvement over a baseline LSTM trained only on raw features. These results highlight the value of embedding nonlinear ecological dynamics prior to temporal modeling, leading to stronger predictive performance and more consistent generalization across random seeds. Our code is available at: https://github.com/aparaselli/HAB-Prediction-Research.**

***Index Terms*—harmful algal blooms, convergent cross mapping, empirical dynamic modeling, long short-term memory, deep learning**

## I. Introduction

Bioluminescent waves in Southern California are visible when dinoflagellates such as Lingulodinium Polyedra bloom in large numbers. Being able to forecast the blooms is crucial due to their toxicity to surrounding sea life. Unfortunately, forecasting these blooms has been a significant challenge due to the vast amount of exogenous variables in the ocean, which introduces chaos into the system. Regardless, modeling Chaotic Nonlinear systems has seen significant research across various domains [1][2] as it has many real-world uses in finance, medicine, cognitive science, ecology, and more. Modern large neural network architectures tend to show failures due to their need for sufficiently large and consistent databases [3] and inability to fully capture the spatial dynamics of the phase space reconstructions of these chaotic systems [4].

## II. Methods

### *A. Dataset and Preprocessing*

Cell count of Lingulodinium Polyedra, total dinoflangellites, total diatoms, total phytoplankton, and Akashiwo Sanguinea alongside ammonium, phaeopigments, phosphate, silicate were aquired from SCCOOS Automated Shore Station [5]. The dataset consists of weekly observations collected over a 16.5-year span, yielding 823 observations. Temperature and Salinity were also taken at depths of 0 meters and 5 meters as a part of the Shore Stations Program [6]. Density at these depths were also calculated using the equation of state for seawater. Finally the bottom up temperature anomaly was calculated in line with past work [7]. Throughout this study we classify a Lingolodinium bloom to occur when the cellcount passes the 95th percentile, which was 54911.850 cells/L. Because there was a 20-week gap with no Lingulodinium polyedra observations between May and November 2021, we split the training data to include only samples prior to May 2021. The final out-of-sample prediction set therefore consists of 163 observations collected between November 2021 and December 2024.

### *B. Empirical Dynamic Modeling*

Empirical Dynamic Modeling (EDM) is a framework for modeling dynamic attractors instead of a set of parametric equations [8]-[12]. We hypothesize that the population of lingulodinium follows a a set of rules governed by a dynamic attractor. EDM leverages Takens’ theorem [13] by reconstructing the state space from time-delayed observations, thereby capturing the system’s underlying attractor. In this study we model the weekly lingulodinium polyedra population through cell counts and concentration data. However since we may lack other important ecological variables The incomplete set of variables can be offset utilizing the fact that dynamic attractors can be reconstructed using lags of multiple timeseries in the system [14][15]. In this study, we utilize S-maps, which leverage a nearest-neighbors approach in the state space reconstruction [8], to forecast Lingulodinium polyedra cell counts. We built and ensemble of EDM models by sampling lag values of environmental variables identified as significant. To determine significant environmental variables, we utilized Convergent Cross Mapping (CCM). Causal influence is determined by CCM and is measured by the ability of Lingulodinium Polyedra to predict past (lagged) values of the candidate variable. Table 1 highlights the lags which we constituted as significant due to them having a rho $>$ 0.1 and convergent p value $<$ 0.2. We chose to use a relatively largely p value due the small size of the dataset.

| **Candidate Variable** | **Prediction Time (weeks)** | **CCM Value** | **Linear Cross Correlation** |
|---|---|---|---|
| Avg Chloro | 0 | 0.969 | 0.955 |
| Avg Phaeo | 0 | 0.947 | 0.900 |
| Total Dinoflagellates | -2 | 0.761 | 0.757 |
| Total Phytoplankton | -1 | 0.672 | 0.542 |
| Total Diatoms | -1 | 0.641 | -0.009 |
| Total Cochlodinium spp | -2 | 0.303 | -0.002 |
| Ammonium | 0 | 0.166 | -0.039 |
| Bottom Density (5m) | -2 | 0.1522 | 0.005 |
| Average Density | -1 | 0.149 | -0.017 |
| Surface Density (0m) | -2 | 0.144 | -0.037 |
| Surface Temp (0m) | -1 | 0.128 | 0.036 |
| Bottom Temp (5m) | -1 | 0.127 | 0.000 |

Table 1: Candidate variables (Yi) listed measured by the ability of Lingoldinium Polyedra to predict Yi. Only candidate variables with CCM scores $>$ 0.1 and convergent p values $<$ 0.2 are listed. Prediction time refers to the optimal lag value through CCM, for example a value of -1 would mean a 1 week lag of Lingulodinium Polyedra performs the best.

We constructed an ensemble of EDM models by varying both the number of lags and the nonlinearity parameter . Specifically, we sampled lag dimensions between 3 and 10, and for each lag dimension we generated 1,000 random samples of lags for each of the following values 1, 5, 9, 15, 25, 45. This resulted in a total of $(103 + 1)1,0006 = 48,000$ models. For the "significant" environmental variables, we allowed lags ranging from 0 to 2 weeks. In all cases, the first coordinate of each EDM model was fixed as the unlagged Lingulodinium polyedra cell count. After ranking the models by their individual cross-mapping score (), we examined which lags most frequently appeared in the top 5% of models (Table 1), providing insight into the interpretability of the system.

### C. Hybrid EDM–LSTM Model

We first generate empirical dynamic modeling (EDM) embeddings of the target series and associated covariates. These embeddings are standardized using train-only means and standard deviations:

$$\mathbf{x}'_t = \frac{\mathbf{x}_t - \boldsymbol{\mu}_{\text{train}}}{\boldsymbol{\sigma}_{\text{train}}}.$$

Sliding windows of length $L$ (hyperparameter `seq_len`) are constructed, each paired with a binary label $y_{t+L} \in \{0, 1\}$ (bloom vs. no bloom).

Before temporal modeling, each feature is reweighted by a learnable gate. Let $\mathbf{z} \in \mathbb{R}^d$ be gate logits. The gate vector is

$$\mathbf{g} = \sigma\left(\frac{\mathbf{z}}{\tau}\right), \qquad \tilde{\mathbf{X}}_t = \mathbf{X}_t \odot \mathbf{g},$$

where $\tau$ is a temperature parameter and $\odot$ denotes channelwise multiplication. Initialization is based on feature-level ROC–AUC scores, and regularization encourages sparsity and near-binarity.

The gated sequence $\tilde{\mathbf{X}}_t$ is passed through:

1) two Temporal Convolutional Network (TCN) layers,
2) a multi-head self-attention block,
3) a stacked LSTM with hidden size $H$ and $n$ layers.

The final hidden representation $\mathbf{h}_T$ is mapped to class probabilities:

$$\hat{\mathbf{y}} = \text{softmax}(W\mathbf{h}_T + \mathbf{b}),$$

where $\hat{\mathbf{y}} \in \mathbb{R}^2$ represents the predicted probabilities of "no bloom" and "bloom."

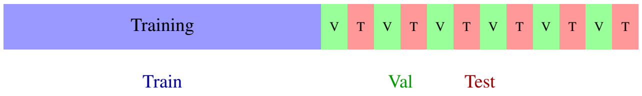
### D. Training

The model is then trained using class-weighted cross-entropy loss,

$$\mathcal{L} = -\sum_i w_{y_i} \log \hat{y}_{i,y_i} + \mathcal{R}_{\text{gate}},$$

where $w_{y_i}$ are class weights, and $\mathcal{R}_{\text{gate}}$ is the gate regularization term. Optimization is performed with Adam, gradient clipping, and exponential moving average (EMA) of parameters. The best snapshot is selected by validation ROC–AUC.

After constructing training windows from the first $N_{\text{train}}$ observations, the remaining post-train windows were divided using an interleave split. In this scheme, the sequence of post-train windows was ordered chronologically, and then assigned alternately to validation and test. This ensured that both sets spanned the same temporal period and contained similar seasonal and distributional variability, while remaining disjoint.



# III. Results

## A. Single–Run ROC Curves

Fig. 1 and Fig. 2 shows representative ROC curves for both the baseline LSTM (trained directly on raw standardized time–series features) and the Hybrid EDM–LSTM (trained on EDM embeddings with gating, convolutional, and recurrent layers). The baseline model reaches moderate performance (Validation AUC ≈ 0.58, Test AUC ≈ 0.64), while the hybrid model achieves higher discriminative ability (Validation AUC ≈ 0.69, Test AUC ≈ 0.75). These single–run curves illustrate the improvement gained from embedding ecological dynamics before temporal modeling.

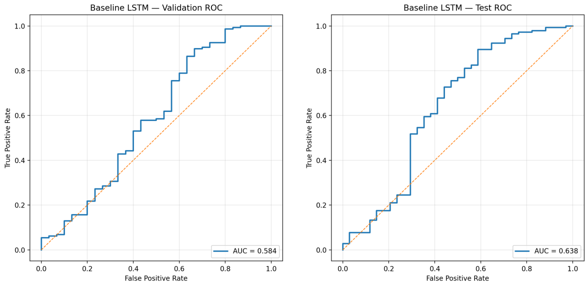


Fig. 1: Baseline LSTM ROC Val/Test curves

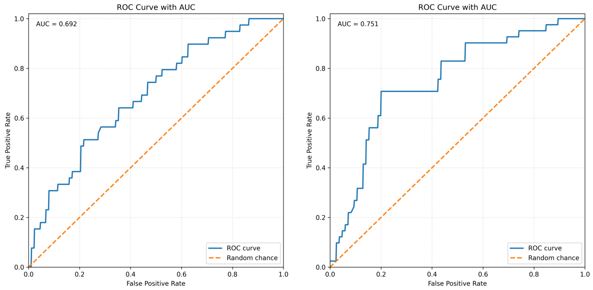


Fig. 2: Hybrid EDM–LSTM ROC Val/Test curves

## B. Multi–Run Evaluation

To assess stability, both models were trained across 60 random seeds using the same interleave split protocol.

- **Distribution of Test AUCs:** As shown in Fig. 3, the baseline LSTM has a median Test AUC around 0.59, while the Hybrid EDM–LSTM centers near 0.71 with reduced variance, indicating both stronger and more consistent generalization.
- **Validation vs. Test Agreement:** Fig. 4 plots Validation AUC against Test AUC across runs. The baseline (blue) points scatter more widely, while the hybrid (red) points cluster closer to the diagonal with a clear upward shift. This suggests that EDM embeddings improve the reliability of validation as a proxy for out–of–sample performance.

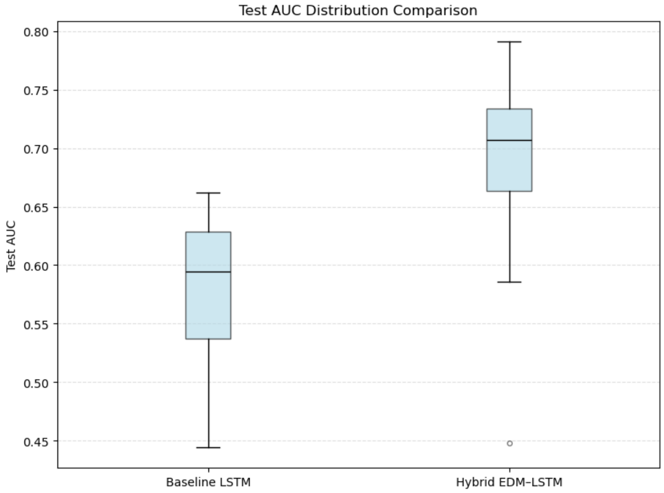


Fig. 3: Distribution of test AUCs across repeated runs. The Hybrid EDM–LSTM achieves higher medians and reduced variance compared to the baseline LSTM.

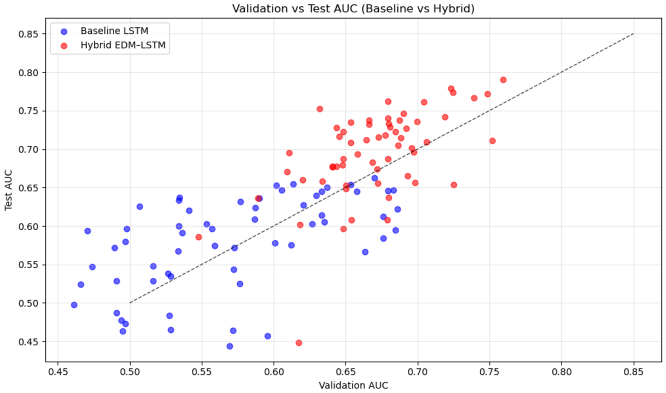


Fig. 4: Validation vs. test AUC across runs. Hybrid EDM–LSTM points (red) lie closer to the diagonal, showing stronger validation–test agreement than the baseline LSTM (blue).

## C. Summary

Overall, results show that the Hybrid EDM–LSTM outperforms a baseline LSTM trained directly on raw ecological covariates. Incorporating EDM embeddings with gating, convolutional, and recurrent layers provides both higher predictive accuracy and more stable generalization across random seeds.

## References

[1] arXiv:2406.11993, “Delay embeddings,” 2024. [Online]. Available: https://arxiv.org/abs/2406.11993 (accessed Aug. 28, 2025).

[2] arXiv:2409.15771, “Chaotic Chrono,” 2024. [Online]. Available: https://arxiv.org/abs/2409.15771 (accessed Aug. 28, 2025).

[3] arXiv:2403.07815, “Chronos,” 2024. [Online]. Available: https://arxiv.org/pdf/2403.07815 (accessed Aug. 28, 2025).

[4] IEEE Xplore Doc. 9183934, “Hybrid neural network (details per paper title),” 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9183934 (accessed Aug. 28, 2025).

[5] M. Carter, “Dataset on CalOOS.” [Online]. Available: https://data.caloos.org/ (accessed Aug. 28, 2025).

[6] Scripps Institution of Oceanography, “Shore Stations Program: Index Data.” [Online]. Available: http://shorestation.ucsd.edu/data/index$_{d}ata.html(accessedAug.28, 2025)$.

[7] Ecology, “Red tide EDM forecast,” [Journal article]. [Online]. Available: https://www.esajournals.onlinelibrary.wiley.com/doi/10.1002/ecy.1804 (accessed Aug. 28, 2025).

[8] Nature, 1990, “Empirical Dynamic Modeling (EDM) (classic paper).” [Online]. Available: https://www.nature.com/articles/344734a0 (accessed Aug. 28, 2025).

[9] *Science*, 2012, “Empirical modeling.” [Online]. Available: https://www.science.org/doi/10.1126/science.1227079 (accessed Aug. 28, 2025).

[10] PubMed, 2015, “EDM 2015.” [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/25829536/full-view-affiliation-1 (accessed Aug. 28, 2025).

[11] Scientific Reports, 2015, “EDM 2015b.” [Online]. Available: https://www.nature.com/articles/srep14750 (accessed Aug. 28, 2025).

[12] Proc. R. Soc. B, 2016, “EDM 2016.” [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rspb.2015.2258 (accessed Aug. 28, 2025).

[13] F. Takens, “Detecting strange attractors in turbulence,” in *Dynamical Systems and Turbulence*, Warwick 1980. Lecture Notes in Math., vol. 898, Springer, 1981. [Online]. Available: https://www.math.ucdavis.edu/ saito/data/synchrosqueezing/takens$_{d}etect-strange-attractors-turbulence.pdf(accessedAug.28, 2025)$.

[14] S. E. A. Sauer, J. A. Yorke, and M. Casdagli, “Embedology,” *J. Stat. Phys.*, 1991. [Online]. Available: https://link.springer.com/article/10.1007/BF01053745 (accessed Aug. 28, 2025).

[15] PLOS ONE, 2011, “(Paper related to EDM / forecasting),” [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0018295 (accessed Aug. 28, 2025).