Jordan Reichhardt

Comparing the Predictions of Chlorophyll Concentrations: A Thesis Introduction

Importance of chlorophyll predictions

- Oxygen depletion:
  - Hypoxia due to algae decomposition causes fish mortality, and ecosystem stress

- Toxin production:
  - Common toxin produced are dominic acid and paralytic shellfish toxins
  - The results are marine mammal illness, shellfish harvest collapse, and neurological illness in cases were toxic shellfish are consumed
- Economic effects:
  - HABs put harmful stress on food webs and thus fisheries, beach access and public health

https://smartwatermagazine.com/news/smart-water-magazine/harmful-algal-blooms-threaten-californias-waters

Widely Used Methods for chlorophyll concentrations measurements

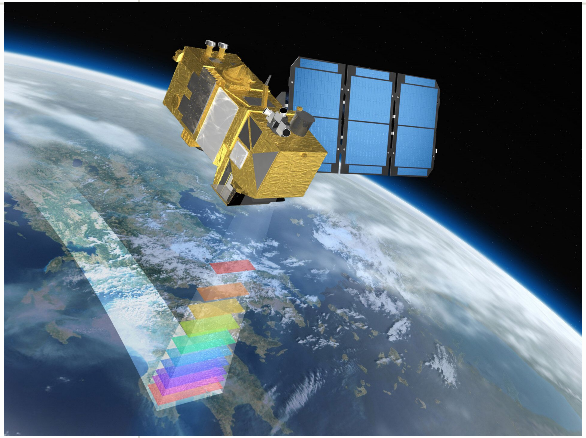
01

In-situ chlorophyll measurements

Fluorometers are used to directly measure the presence of phytoplankton for the given volume of water

https://santamariatimes.com/news/local/cal-poly-hosts-annual-open-house-on-its-pier/article_c2edb8f7-9cb1-5911-89ae-4484d9336a67.html

02



https://www.esa.int/Enabling_Support/Operations/Sentinel_mission_control_taking_shape

Satellite predictions

Satellites measure the amount of sunlight reflected from the ocean's surface and apply a ration of blue/ green wavelength relations to calculate the chlorophyll concentrations

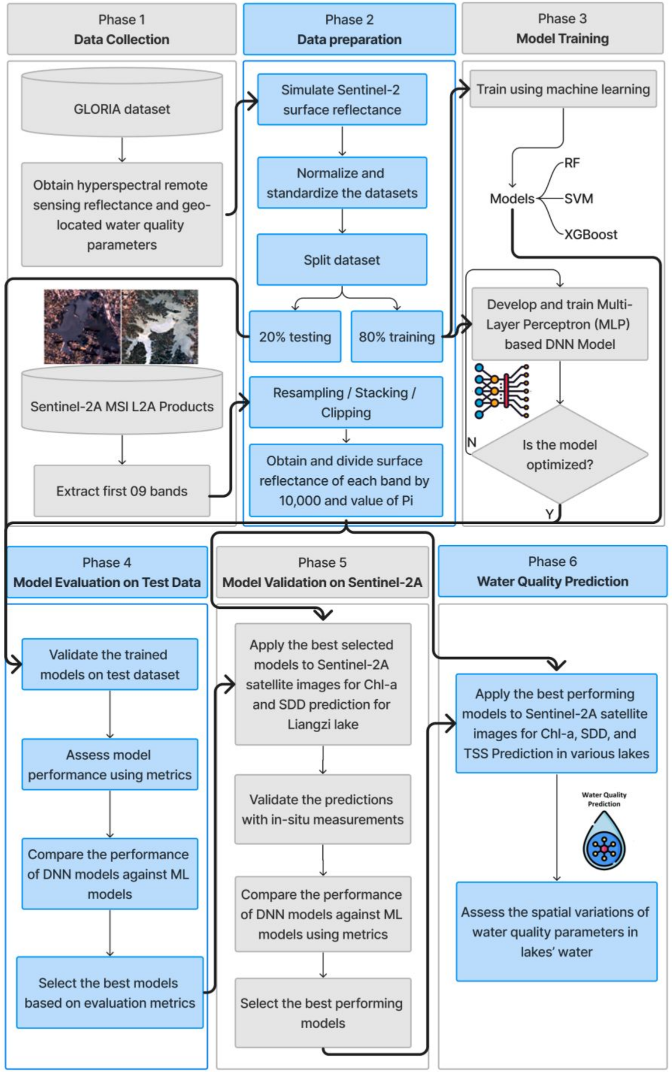
Limitations of Current Methods for chlorophyll measurements

01 In-situ chlorophyll measurements

- Require expensive equipment and extensive installation process
- Not practical for many areas of interest

02 Satellite predictions

- Temporal limitations (1-7 day gaps)
- Low accuracy in coastal regions
- Land pixel interference



Foundational Research:

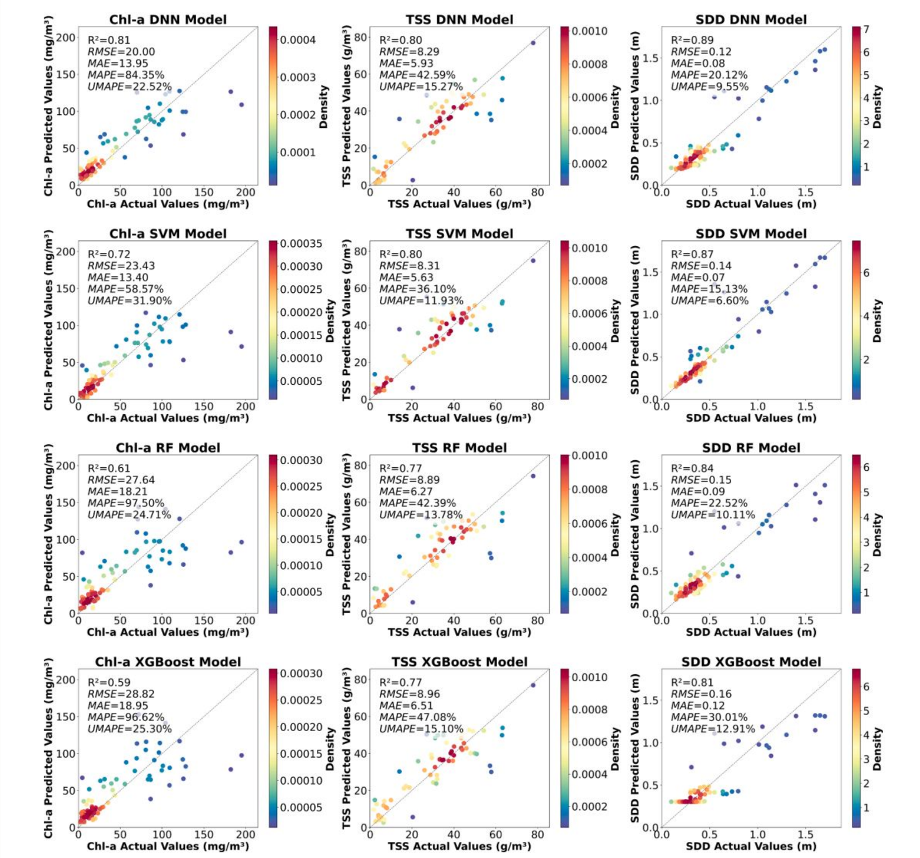
“Deep learning for water quality multivariate assessment in inland water across China”

Source: Ali et al., 2023 [4]

- Ali et al. use Sentinel 2 inputs and various ML algorithms to predict chlorophyll in multiple inland lakes in China
- Trained ML algorithms solely off of Satellite data compared to measured values
- Larger chlorophyll range and larger RMSE

Foundational Research:

“Deep learning for water quality multivariate assessment in inland water across China”

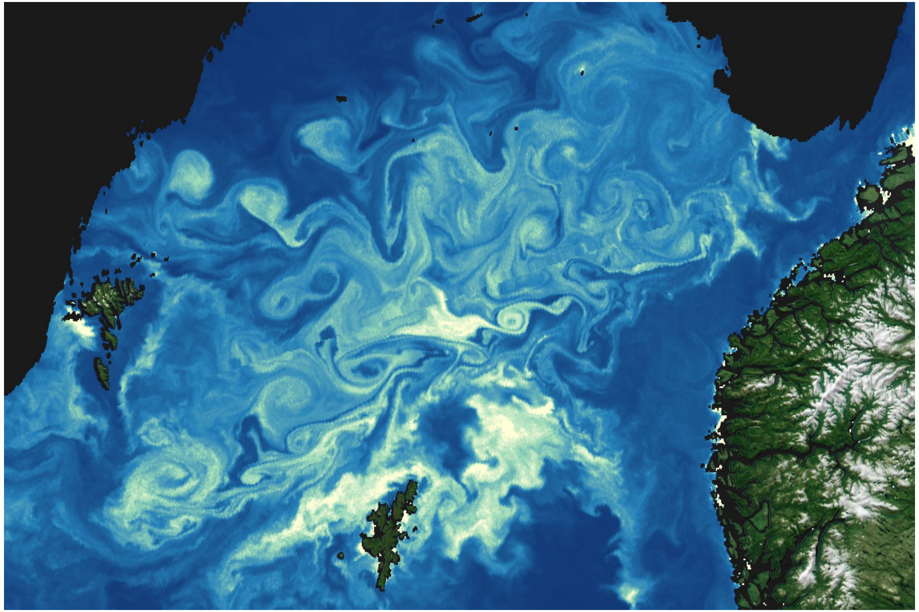


Source: Ali et al., 2023 [4]

- Ali et al. use Sentinel 2 inputs and various ML algorithms to predict chlorophyll in multiple inland lakes in China
- Trained ML algorithms solely off of Satellite data compared to measured values
- Larger chlorophyll range and larger RMSE

Problem Statement

- Question:
- Can Machine Learning models combine environmental observations to enhance predictions of chlorophyll over satellite retrieval alone?



Area of Study

Cal Poly Pier (-120.7425 W, 35.1675 S) - NOAA station CPXC1

Data Source	Retrieval Location	Frequency of Retrieval	Data Role
Environmental Data	CPXC1 — Cal Poly Pier Monitoring Station	30-minute measurements	Input Feature
Swell Data	NOAA Station 46215 — Diablo Canyon Offshore Buoy	30-minute aligned observations	Input Feature
Satellite Data	Sentinel-3 OLCI Satellite Retrieval	(1-7 days)→ aligned to 30-minute timestamps using the fill-forward technique	Input Feature
Ground Truth Data	Cal Poly Pier Profiler (In Situ Measurement)	30-minute measurements	Ground Truth / Prediction Target

Environmental Forcing Features

Cal Poly Pier (-120.7425 W, 35.1675 S) - NOAA station CPXC1

Variable	Description	Units
Wind Speed	Average wind speed	m/s
Wind Gust	Maximum wind speed	m/s
Wind Direction	Wind direction	degrees
Air Temperature	Atmospheric temperature	°C
Relative Humidity	Atmospheric moisture	%
Atmospheric Pressure	Surface pressure	hPa
Solar Radiation	Incoming solar energy	W/m^2
Net Atmospheric Radiation	Net radiative exchange	W/m^2
Rainfall	Precipitation accumulation	mm
Dew Point	Moisture condensation temperature	°C

Swell Forcing Features

NOAA station 46215

https://www.ndbc.noaa.gov/station_page.php?station=46025

Variable	Description	Units
Mean Wave Period	Average interval between successive waves	s
Peak (Dominant) Wave Period	Wave period matched to the maximum energy in the wave spectrum	s
Significant Wave Height	Average height of the highest one-third of observed waves	m
Wave Direction	The direction in which the waves are propagating	degrees

Satellite Forcing Features

Sentinel 3 retrievals

Variable	Description	Units
Open Ocean Chlorophyll (Filled Forward)	Sentinel-3 open ocean chlorophyll-a estimate (OC4Me algorithm). Missing satellite observations are propagated forward.	µg/L
Open Ocean Source Time	Timestamp of the original Sentinel-3 open ocean chlorophyll observation used to populate the filled-forward value.	UTC datetime
Open Ocean Data Age	Time elapsed between the model timestamp and the original open ocean satellite observation. Used to indicate how stale the filled-forward value is.	hr
Complex Waters Chlorophyll (Filled Forward)	Sentinel-3 optically complex waters chlorophyll-a estimate (C2RCC algorithm). Missing observations are filled forward using the most recent valid retrieval.	µg/L
Complex Waters Source Time	Timestamp of the original Sentinel-3 complex waters chlorophyll observation used to populate the filled-forward value.	UTC datetime
Complex Waters Data Age	Time elapsed between the model timestamp and the original complex waters satellite observation. Indicates the confidence in the temporal alignment of the satellite estimate.	hr

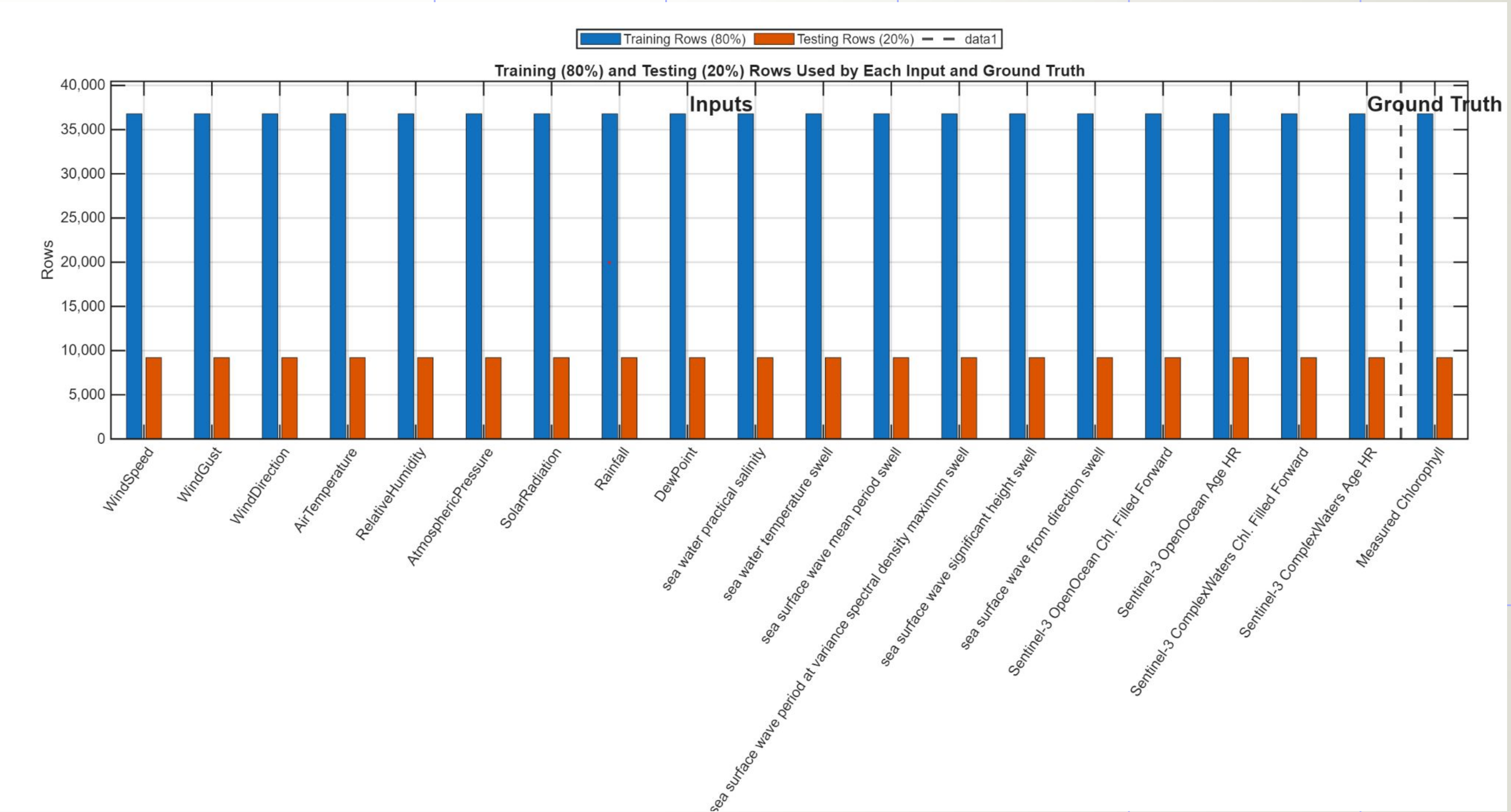
https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-3

Chlorophyll Target Values

Cal Poly Pier (-120.7425 W, 35.1675 S) - NOAA station CPXC1

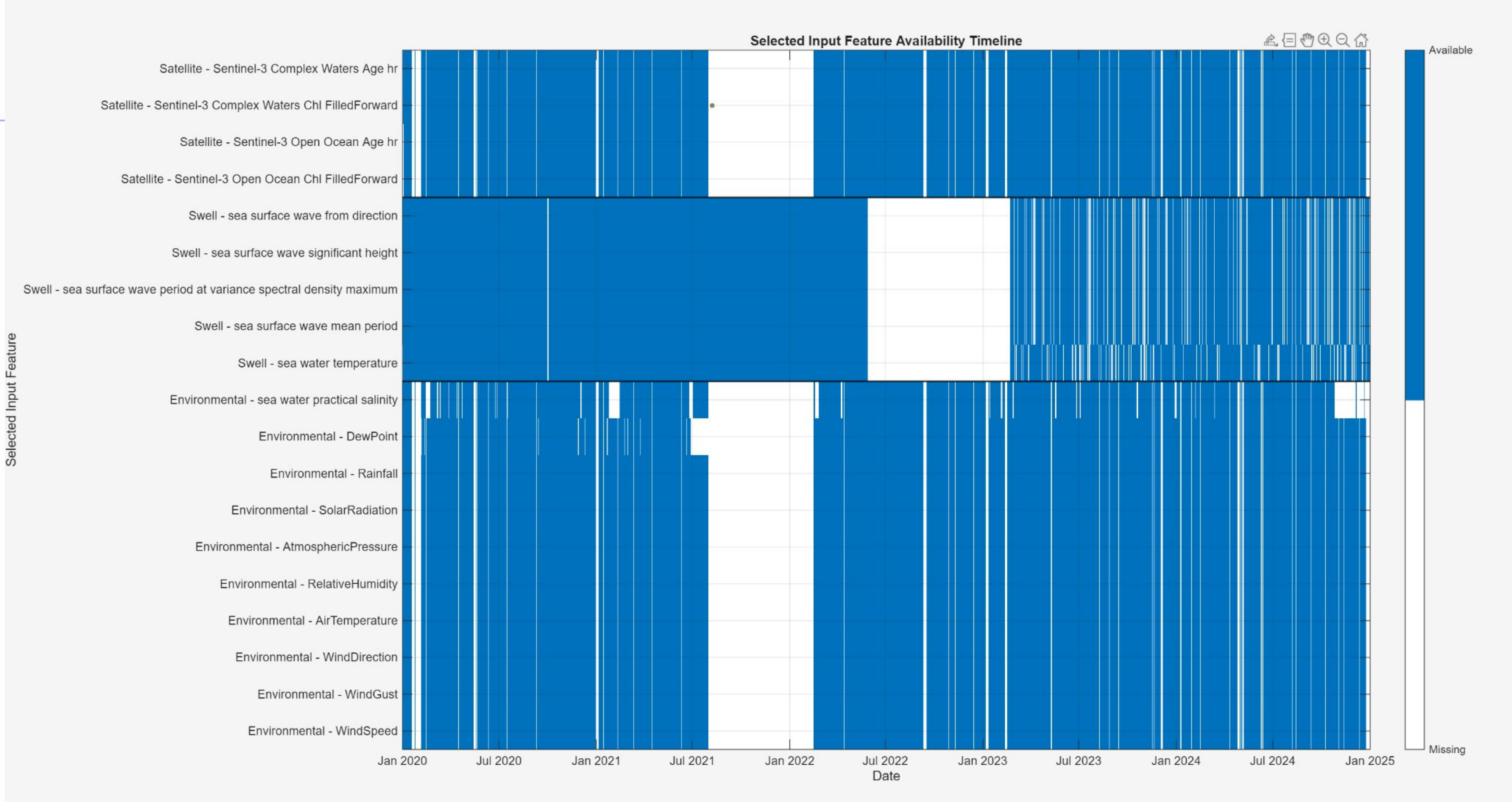
Variable	Description	Units
Chlorophyll-a	Biomass proxy for phytoplankton abundance	µg/L

Datasets



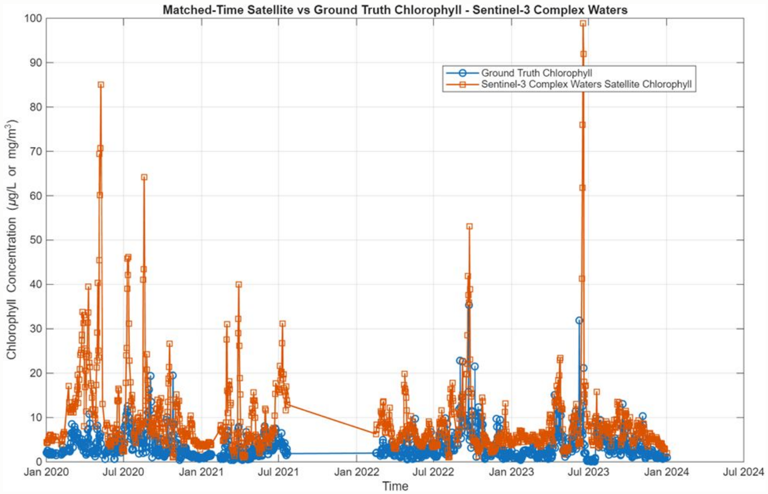
Training Rows (80%)
Testing Rows (20%)
data1
Training (80%) and Testing (20%) Rows Used by Each Input and Ground Truth
Inputs
Ground Truth
Rows
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
WindSpeed
WindGust
WindDirection
AirTemperature
RelativeHumidity
AtmosphericPressure
SolarRadiation
Rainfall
DewPoint
sea water practical salinity
sea water temperature swell
sea surface wave mean period swell
sea surface wave period at variance spectral density maximum swell
sea surface wave significant height swell
sea surface wave from direction swell
Sentinel-3 OpenOcean Chl. Filled Forward
Sentinel-3 OpenOcean Age HR
Sentinel-3 ComplexWaters Chl. Filled Forward
Sentinel-3 ComplexWaters Age HR
Measured Chlorophyll

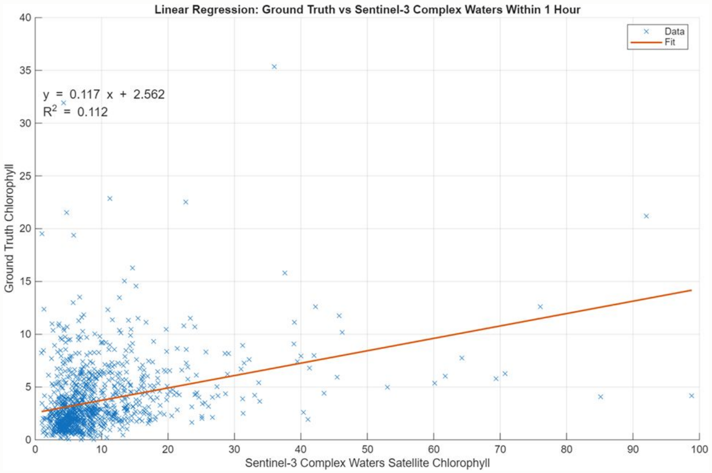
Datasets


Selected Input Feature Availability Timeline
Selected Input Feature
Satellite - Sentinel-3 Complex Waters Age hr
Satellite - Sentinel-3 Complex Waters Chl FilledForward
Satellite - Sentinel-3 Open Ocean Age hr
Satellite - Sentinel-3 Open Ocean Chl FilledForward
Swell - sea surface wave from direction
Swell - sea surface wave significant height
Swell - sea surface wave period at variance spectral density maximum
Swell - sea surface wave mean period
Swell - sea water temperature
Environmental - sea water practical salinity
Environmental - DewPoint
Environmental - Rainfall
Environmental - SolarRadiation
Environmental - AtmosphericPressure
Environmental - RelativeHumidity
Environmental - AirTemperature
Environmental - WindDirection
Environmental - WindGust
Environmental - WindSpeed
Jan 2020
Jul 2020
Jan 2021
Jul 2021
Jan 2022
Jul 2022
Jan 2023
Jul 2023
Jan 2024
Jul 2024
Jan 2025
Date
Available
Missing

Analysis of Current Satellite Predictions

Location of predictions and ground truth measurements: Cal Poly Pier (-120.7425 W, 35.1675 S)

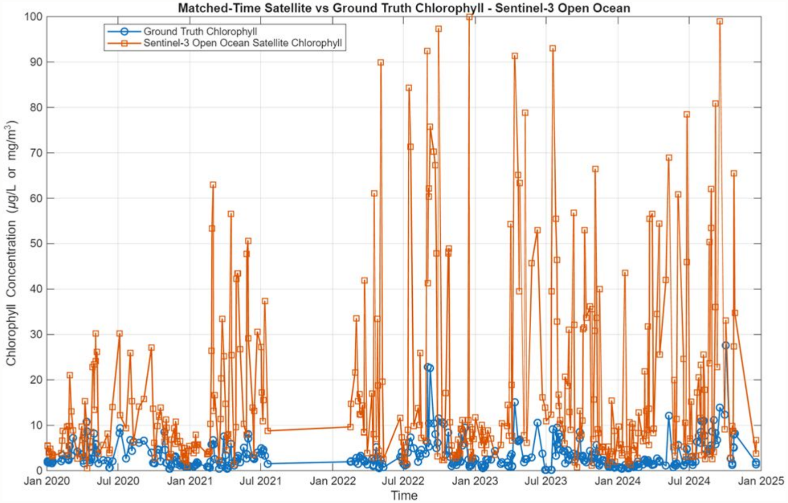


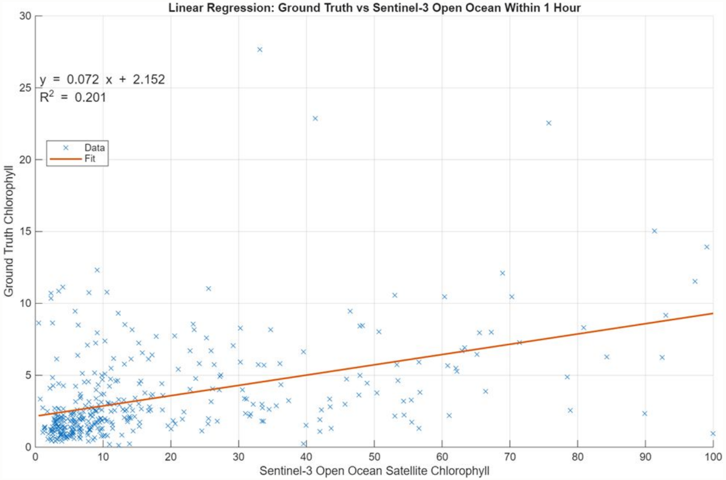


Time aligned comparison of 300 chlorophyll satellite predictions compared to ground truth values

Analysis of Current Satellite Predictions

Location of predictions and ground truth measurements: Cal Poly Pier (-120.7425 W, 35.1675 S)





Time aligned comparison of 300 chlorophyll satellite predictions compared to ground truth values

ML models used

2. XGBoost

- Builds many consecutive decision trees
- The following tree predicts the residual error from the previous tree
- Inputs split by decision thresholds at the tree nodes
- Training creates a new tree and does not continuously modify the current network

3. Feedforward Neural Network

- Learns continuous non-linear relationships using neurons and weights
- Inputs processed simultaneously
- Learns via back propagation prediction error which allows for repeated weight updates
- Training keeps updating the same network

XGBoost Structure Code

```
#Training/testing

X = model_df[features] #env., swell, and sattelite data
y = model_df[target] #chlorophyll ground truth data

X_train, X_test, y_train, y_test, time_train, time_test = train_test_split(
    X,
    y,
    model_df[time_col],
    test_size=0.2,
    random_state=31
)
```

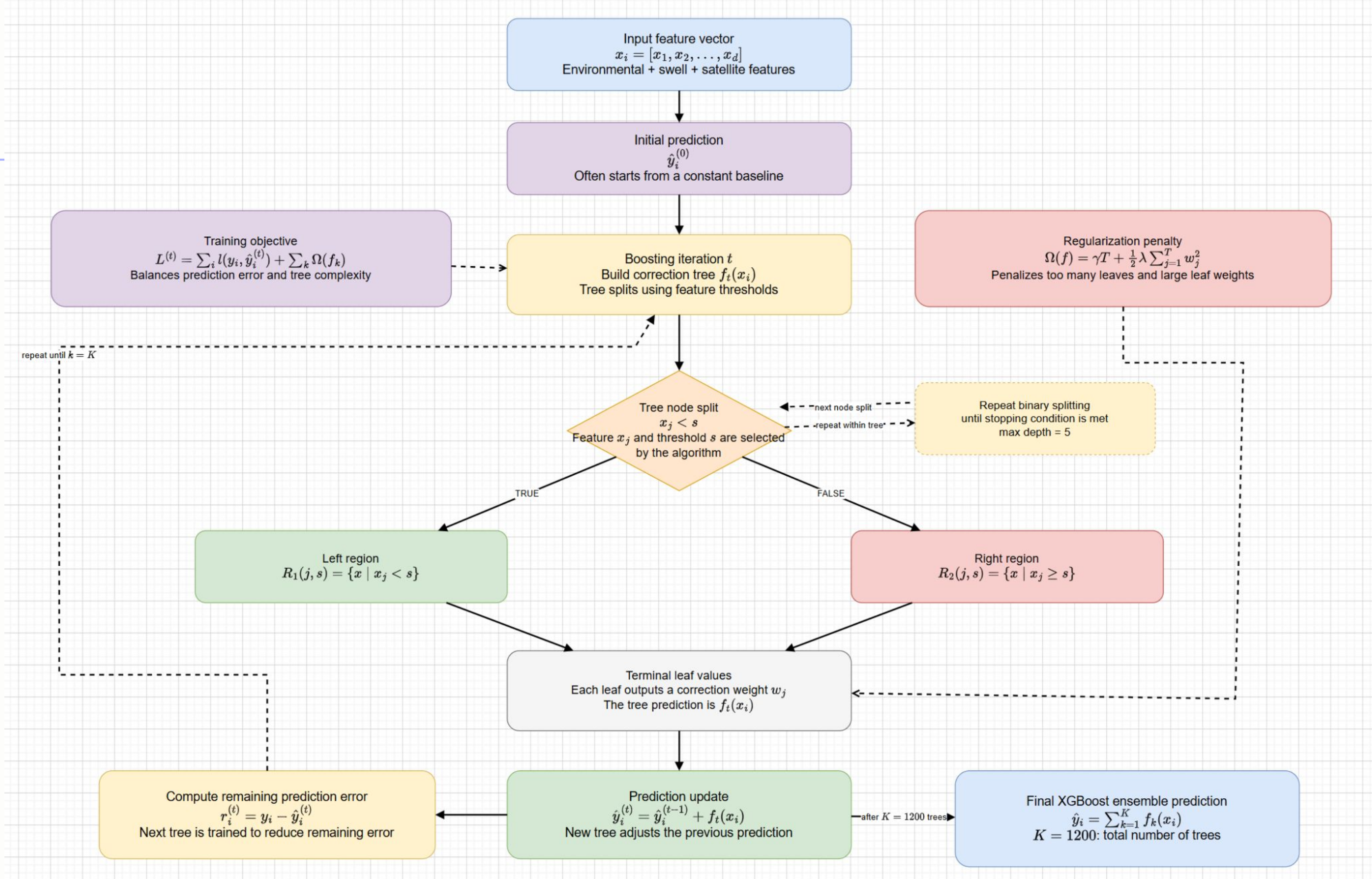
```
#Train model

model = XGBRegressor(
    n_estimators=1200, #of trees
    max_depth=5, #max depth of branches on the tree
    subsample=0.7,#fraction of data rows per tree
    colsample_bytree=0.7, #fraction of features per tree
    reg_alpha=0.5, #more weight to simpler trees
    reg_lambda=2.0, #less weight for large corrections
    min_child_weight=4, #min rows to split so we dont split small env. groups
    objective="reg:squarederror", #loss function: L = (y - y_pred)^2
    n_jobs=-1, #use all CPU cores
    random_state=31
)

model.fit(X_train, y_train)
```

```
#predict using the trained model
y_pred = model.predict(X_test)
y_train_pred = model.predict(X_train)
```

XGBoost

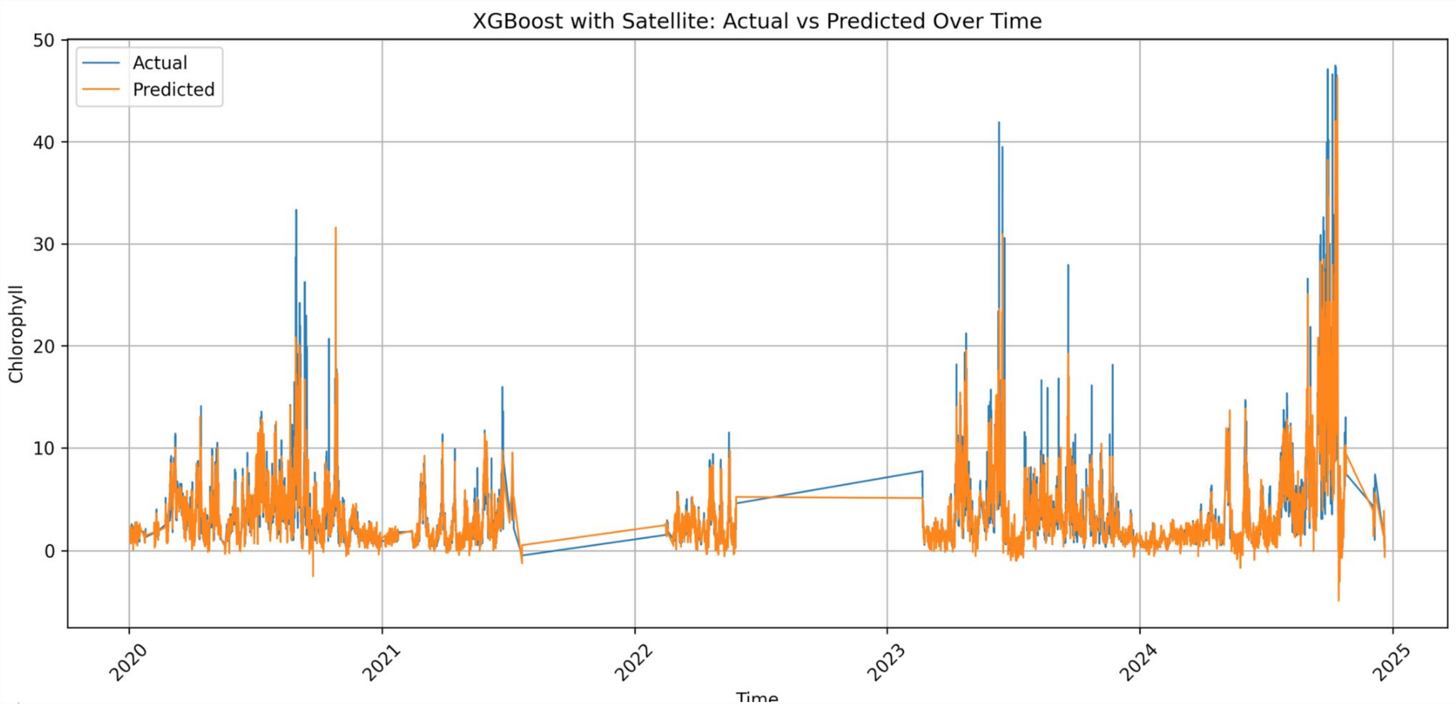
Step	Value	Function	Equation
Input layer	19 inputs	Environmental + swell + satellite inputs	$x = [x_1, x_2, \dots, x_{19}]$
Tree ensemble	1200 trees	Sequentially improve prediction	$F(x) = \sum_{i=1}^{1200} \eta T_i(x)$
Single tree depth	max depth = 5	Maximum of 5 decision levels per tree	Leaves $\leq 2^5 = 32$
Tree split	Binary decision	Split data into environmental regions	$x_j < threshold$
Residual update	Error correction	Next tree predicts remaining error	$r = y - \hat{y}$
Learning rate	$\eta = 0.02$	Controls contribution of each tree	$\hat{y}_k = \hat{y}_{k-1} + \eta T_k(x)$
Subsample	0.7	Uses 70% of rows per tree	Random training subset
Feature sample	0.7	Uses 70% of inputs per tree	Random feature subset
Final output	1 prediction	Combined chlorophyll estimate	$\hat{y} = F(x)$

XGBoost



Input feature vector
$x_i = [x_1, x_2, \dots, x_d]$
Environmental + swell + satellite features
Initial prediction
$\hat{y}_i^{(0)}$
Often starts from a constant baseline
Training objective
$L^{(t)} = \sum_i l(y_i, \hat{y}_i^{(t)}) + \sum_k \Omega(f_k)$
Balances prediction error and tree complexity
Boosting iteration t
Build correction tree $f_t(x_i)$
Tree splits using feature thresholds
Regularization penalty
$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$
Penalizes too many leaves and large leaf weights
repeat until $k = K$
Tree node split
$x_j < s$
Feature x_j and threshold s are selected by the algorithm
next node split
repeat within tree
Repeat binary splitting
until stopping condition is met
max depth = 5
TRUE
FALSE
Left region
$R_1(j, s) = \{x \mid x_j < s\}$
Right region
$R_2(j, s) = \{x \mid x_j \geq s\}$
Terminal leaf values
Each leaf outputs a correction weight w_j
The tree prediction is $f_t(x_i)$
Compute remaining prediction error
$r_i^{(t)} = y_i - \hat{y}_i^{(t)}$
Next tree is trained to reduce remaining error
Prediction update
$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$
New tree adjusts the previous prediction
after $K = 1200$ trees
Final XGBoost ensemble prediction
$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$
$K = 1200$: total number of trees

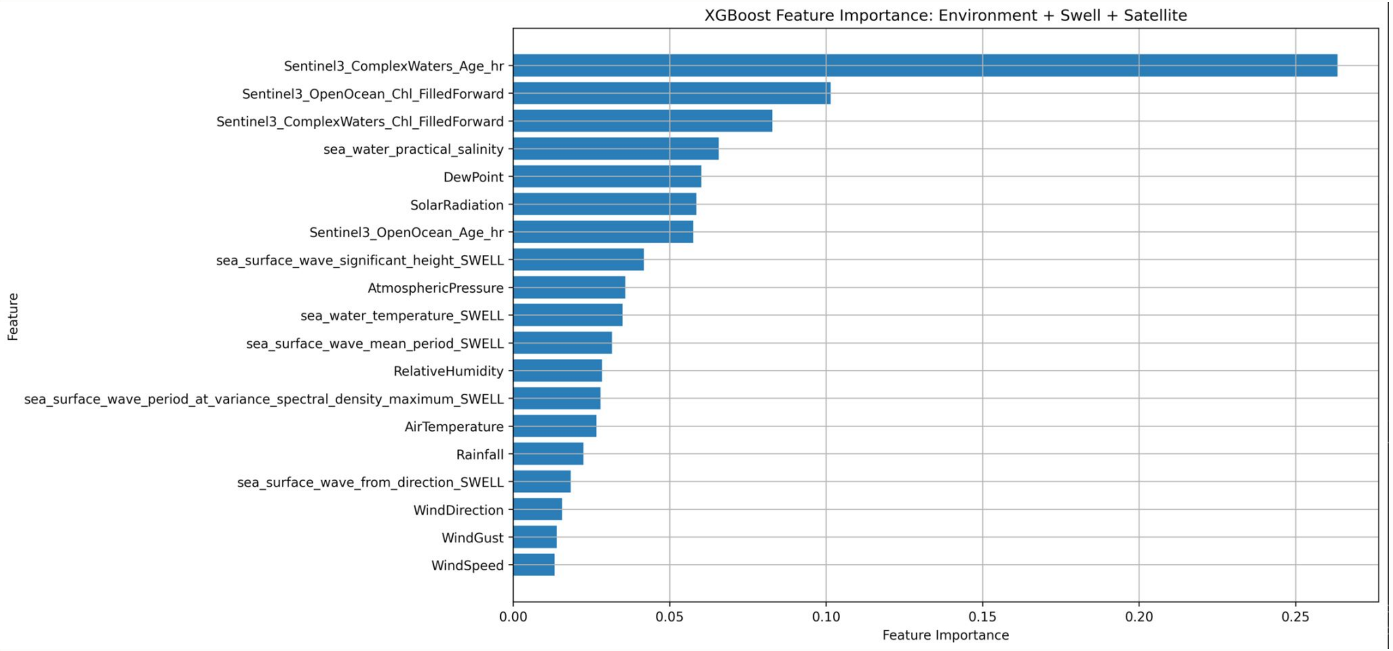
XGBoost Results



XGBoost Results

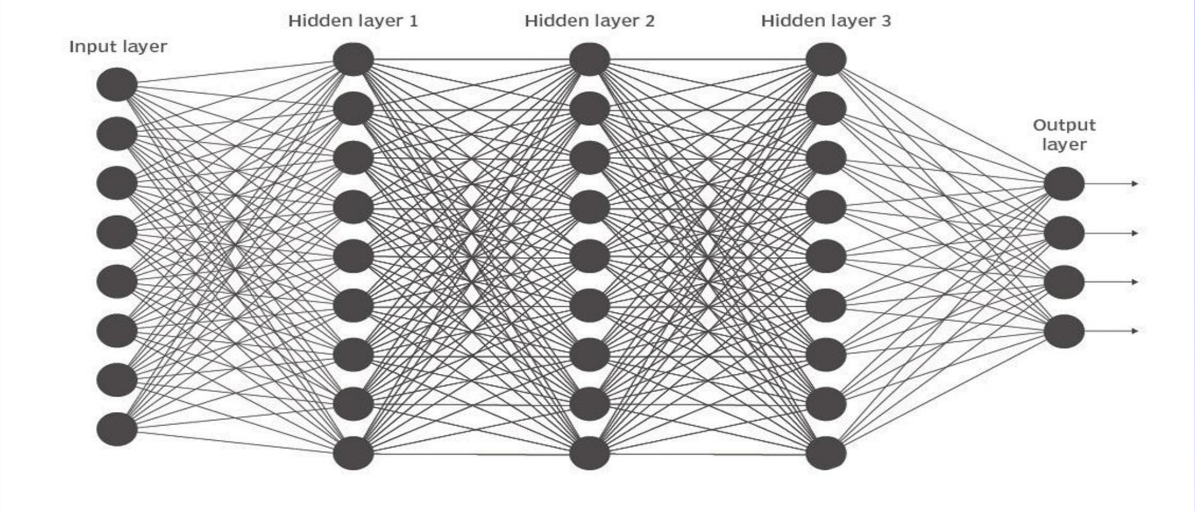
Category	Metric	Value	Meaning
Prediction Accuracy	MAE	**0.794**	Average absolute prediction error
Prediction Accuracy	RMSE	**1.483**	Typical prediction error (model is slightly underpredicting)
Model Fit	R^2	**0.853**	Percent of chlorophyll variability explained on test data
Bias	Bias	**0.002**	Average overprediction or underprediction
Generalization	Train R^2	**0.992**	Model fit on training data
Generalization	Test R^2	**0.853**	Model fit on unseen test data
Overfitting	R^2 Gap	**0.139**	Difference between training and testing performance

XGBoost Results



Feedforward Neural Network

19 128 64 32



https://curiocial.com/convolutional-neural-network-cnn-ai/

FNN code structure

```
# Training/testing split (80:20)
X = model_df[features].values
y = model_df[target].values

X_train, X_test, y_train, y_test, time_train, time_test = train_test_split(
    X,
    y,
    model_df[time_col],
    test_size=0.2,
    random_state=31
)
```

```
# FNN model architecture
model = Sequential([
    Dense(128, activation="relu", input_shape=(X_train_scaled.shape[1],)),
    Dropout(0.20),

    Dense(64, activation="relu"),
    Dropout(0.20),

    Dense(32, activation="relu"),

    Dense(1)
])

model.compile(
    optimizer=tf.keras.optimizers.Adam(learning_rate=0.001),
    loss="mse",
    metrics=["mae"]
)
```

```
#Early end
early_stop = EarlyStopping(
    monitor="val_loss",
    patience=25,
    restore_best_weights=True
)


#Train
history = model.fit(
    X_train_scaled,
    y_train,
    validation_split=0.2,
    epochs=300,
    batch_size=64,
    callbacks=[early_stop],
    verbose=1
)
```

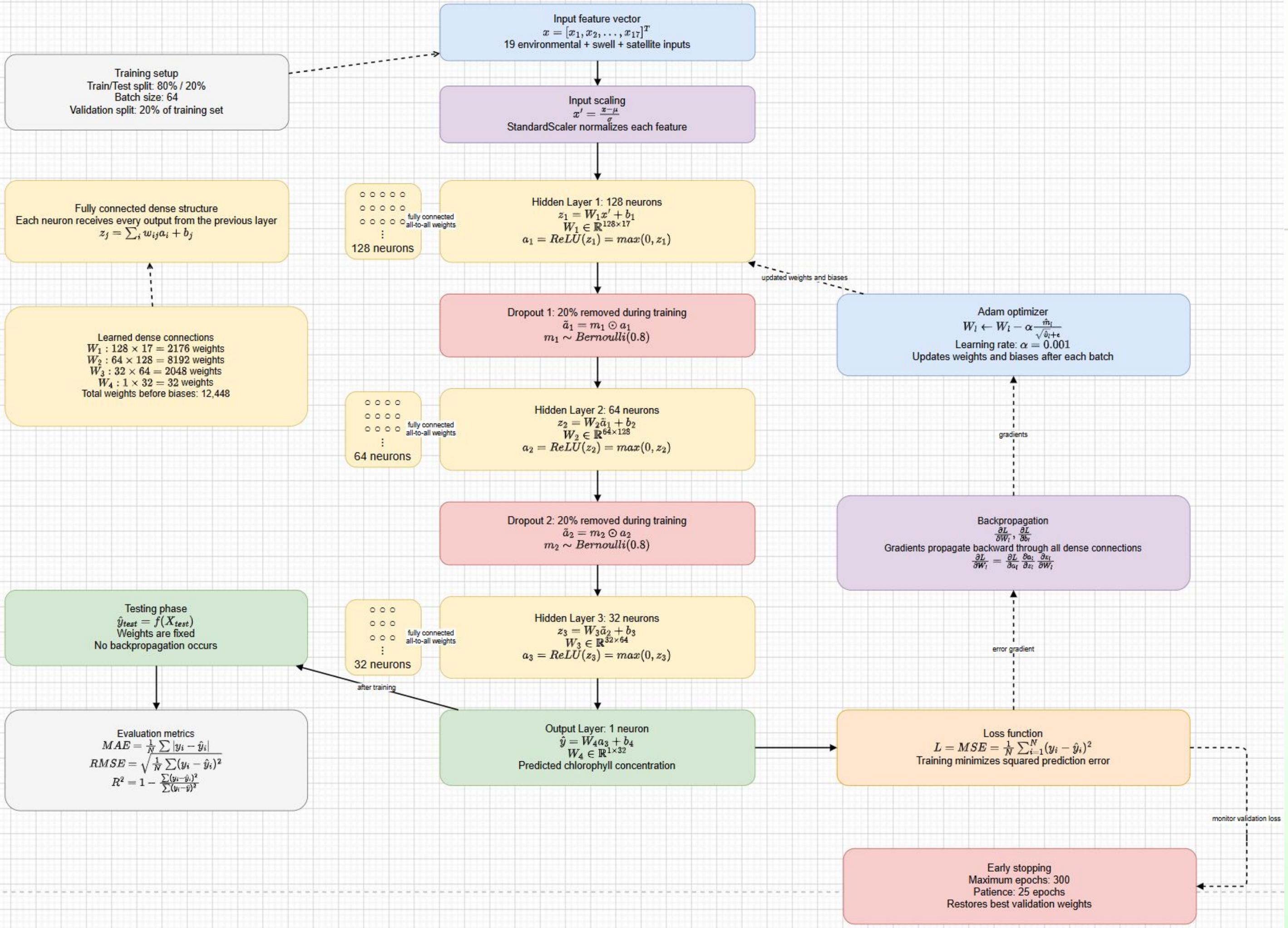
FNN Model breakdown

Steps	Value	Function	Equation
Input layer	19 inputs	env + swell + satellite inputs	$x = [x_1, x_2, \dots, x_{19}]$
Hidden layer 1	128 neurons	First learned feature-combination layer	$z_1 = W_1 x + b_1$ ReLU: $a_1 = \max(0, z_1)$
Dropout 1	0.20	Randomly removes 20% of neurons	$a_{1,\text{drop}} = m \odot a_1$ where $m \sim Bernoulli(0.8)$
Hidden layer 2	64 neurons	Combines outputs from layer 1	$z_2 = W_2 a_1 + b_2$ ReLU: $a_2 = \max(0, z_2)$
Dropout 2	0.20	Removes 20% of active neurons	$a_{2,\text{drop}} = m \odot a_2$
Hidden layer 3	32 neurons	Final learned feature-combination layer	$z_3 = W_3 a_2 + b_3$ ReLU: $a_3 = \max(0, z_3)$
Output layer	1 neuron	Produces chlorophyll prediction	$\hat{y} = W_4 a_3 + b_4$

Variable Definition

- **x:** Environmental input variables
- **W:** Learned feature weights
- **b:** Neuron offset term
- **z:** Weighted input sum
- **a:** Activated output term
- **ReLU:** Nonlinear activation function
- **y^:** Predicted chlorophyll
- **y:** Measured chlorophyll
- **L:** Prediction error (loss)
- **α:** Learning step size

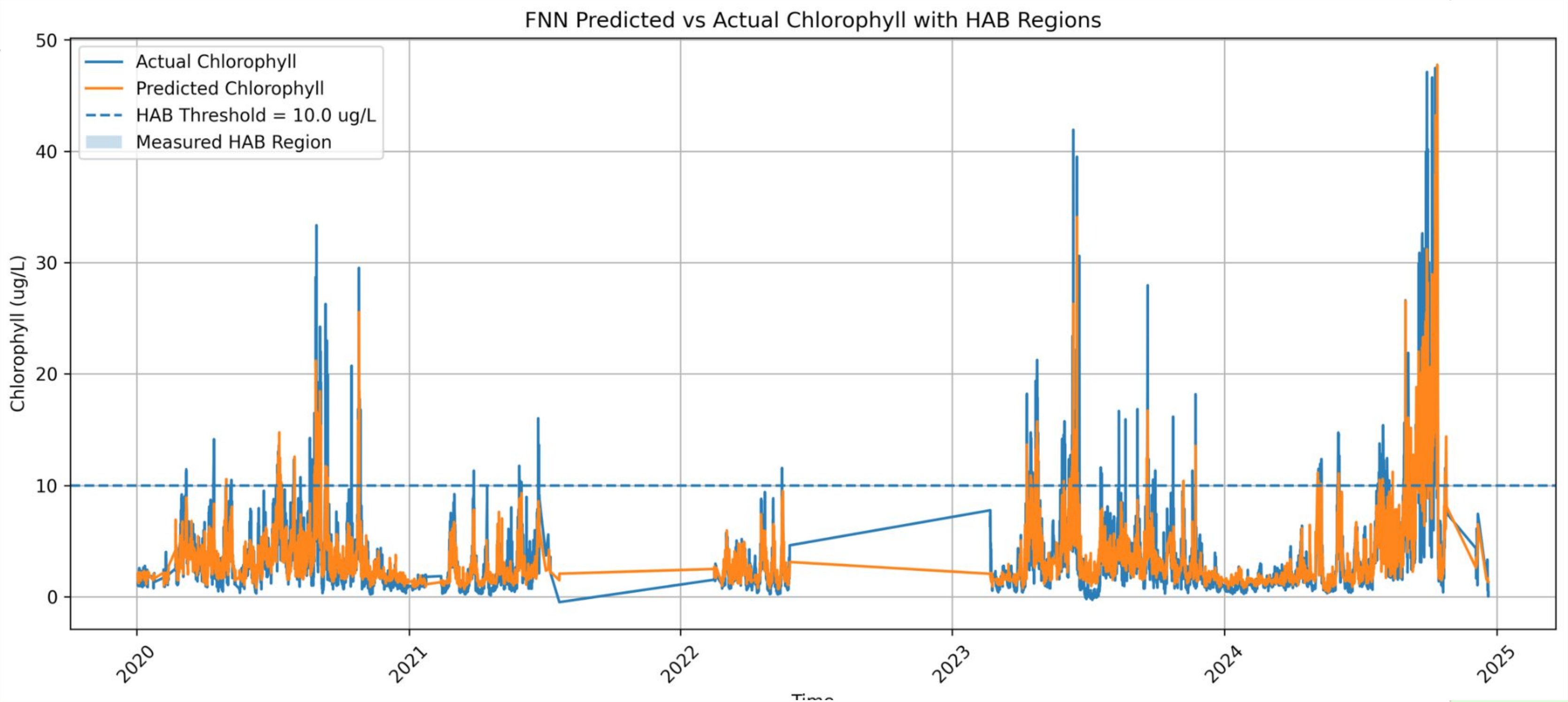
FNN Model breakdown



FNN Training

Category	Metric	Value	Interpretation
Training Duration	Actual Epochs Completed	**207**	Model trained for 207 full passes before stopping
Training Duration	Maximum Epochs Allowed	**300**	Training could run up to 300 epochs
Early Stopping	Stopped Early?	**Yes**	Validation loss stopped improving before 300 epochs
Optimization	Batch Size	**64**	64 samples processed before each weight update
Regularization	Dropout	**20%**	Randomly removes neurons during training to reduce memorization

FNN Results


FNN Predicted vs Actual Chlorophyll with HAB Regions
Actual Chlorophyll
Predicted Chlorophyll
HAB Threshold = 10.0 ug/L
Measured HAB Region
Chlorophyll (ug/L)
50
40
30
20
10
0
2020
2021
2022
2023
2024
2025

FNN Results

Category	Metric	Value	Interpretation
Prediction Error	MAE	1.1544	Average prediction error ≈ 1.15 µg/L chlorophyll
Prediction Error	RMSE	1.9504	Prediction error with larger errors weighted more heavily
Model Fit	R^2	0.7454	The model explains 74.54% of chlorophyll variability
Bias	Bias	−0.1401	Small tendency to underpredict chlorophyll
Training Performance	Train R^2	0.8146	Model fit on training data(80%)
Testing Performance	Test R^2	0.7454	Predictive performance on unseen data (20%)
Overfitting Check	R^2 Gap	0.0692	A small train–test gap means there is a strong generalization

Future work

- Implement a year-year test/ training split to understand the more real predictive power of the inputs ex. (2020-2023 training and 2024 testing)

- Test on a different but similar coastal environment to what was used for training

- Retrieve Sentinel-2 satellite images and create chlorophyll extractions from scratch instead of comparing to Sentinel-3 estimates

Citations

References:

[1] Q. S. Qing, B. Zhang, and X. Zhang, “Improving remote sensing retrieval of water clarity in complex coastal and inland waters with modified absorption estimation and optical water classification,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 102, p. 102323, 2021.

[2] Z.-P. Lee, S. Shang, C. Hu, K. Du, A. Weidemann, W. Hou, J. Lin, and G. Lin, “Secchi disk depth: A new theory and mechanistic model for underwater visibility,” *Remote Sensing of Environment*, vol. 169, pp. 139–150, 2015.

[3] C. Brockmann, R. Doerffer, M. Peters, S. Kerstin, J. Embacher, and A. Ruescas, “Evolution of the C2RCC neural network for Sentinel 2 and 3 for the retrieval of ocean colour products in normal and extreme optically complex waters,” *ESA Living Planet Symposium*, 2016.

[4] A. Ali, G. Zhou, F. P. Antezana Lopez, C. Xu, G. Jing, and Y. Tan, “Deep learning for water quality multivariate assessment in inland water across China,” International Journal of Applied Earth Observation and Geoinformation, vol. 133, p. 104078, 2024.

[5] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.