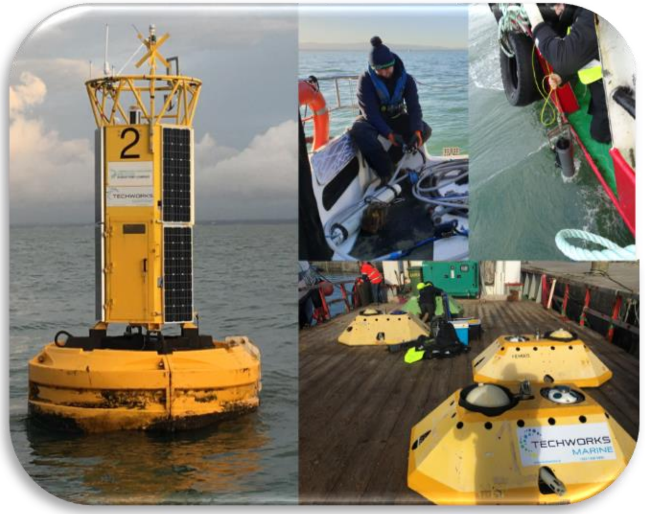
TECHWORKS MARINE

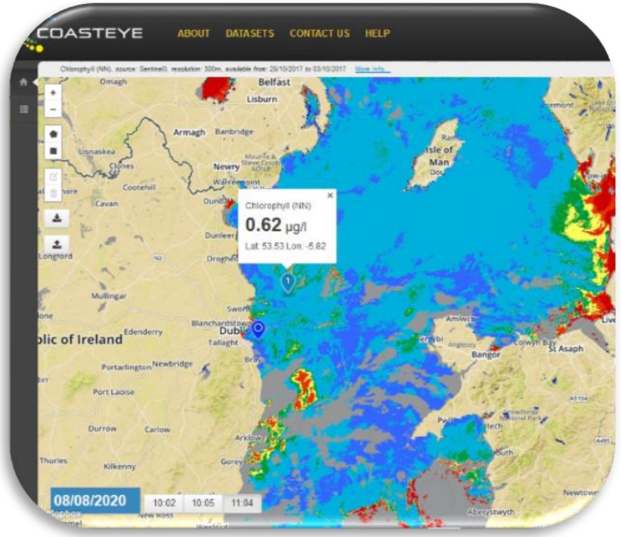
From Space to Sea: A Machine Learning Synthesis of Satellite-Derived and In-Situ Water Quality Analysis in Ireland

Kalliopi Karadima
Earth Observation Scientist

9th Sentinel-3 Validation Team Meeting
30 March-01 April 2026 | ESA-ESRIN

About TechWorks Marine





Integrated Data Buoys

- Real-time enabled
- TMBB data acquisition technology
- Plug and play with a wide range of marine sensors
- 3G, VHF, and Iridium communications

Data Management and Delivery

- Regular field service
- Data telemetry
- Data management including: quality assurance, calibration
- **CoastEye** - live data portal

Value added Products & Services

- Using satellite EO data products (global coverage)
- Validating numerical models
- Hydrographic/Metocean survey provision
- Equipment sales/Rental

Introduction

Problem

- **Coastal zones heavily impacted by human activity**
- **Leads to eutrophication in waters**
- **Chlorophyll-a (Chl-a) = key indicator**

Objective: Validation of satellite measurements of chlorophyll-a in optically complex waters against in-situ measurements, using physics-informed explainable Machine Learning models.

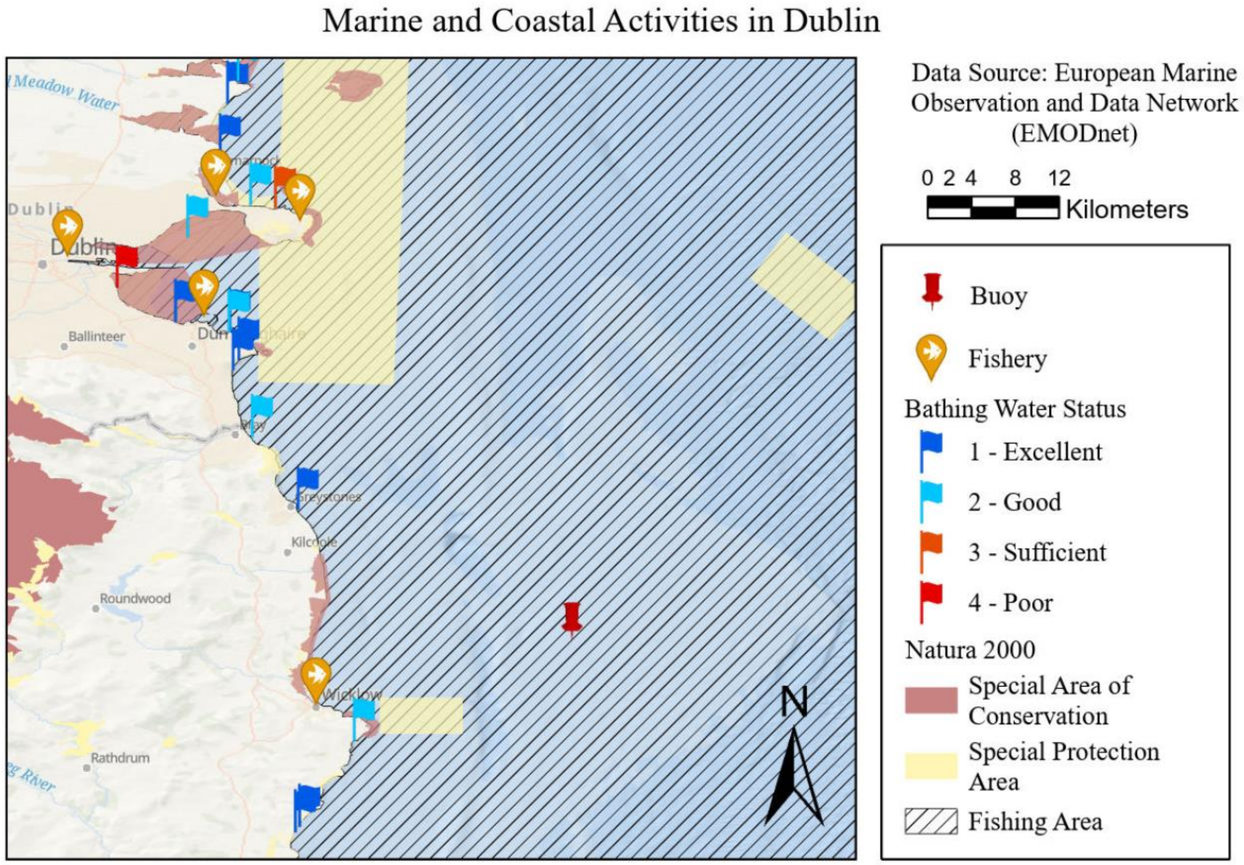


Fig. 1 Overview of marine and coastal activities along the Dublin coast, highlighting fisheries zones, bathymetric contours, and designated bathing spots. Location of the deployed buoy.

In-Situ Data

In-situ data

- Buoy deployed between November 2021 to May 2022
- Location: Irish Sea (53◦2.87586′ N and longitude:5◦5.0025′ W)

Parameters measured (HydroCat)

- Chlorophyll-a
- Conductivity
- Temperature
- Pressure
- Dissolved oxygen
- pH
- Turbidity

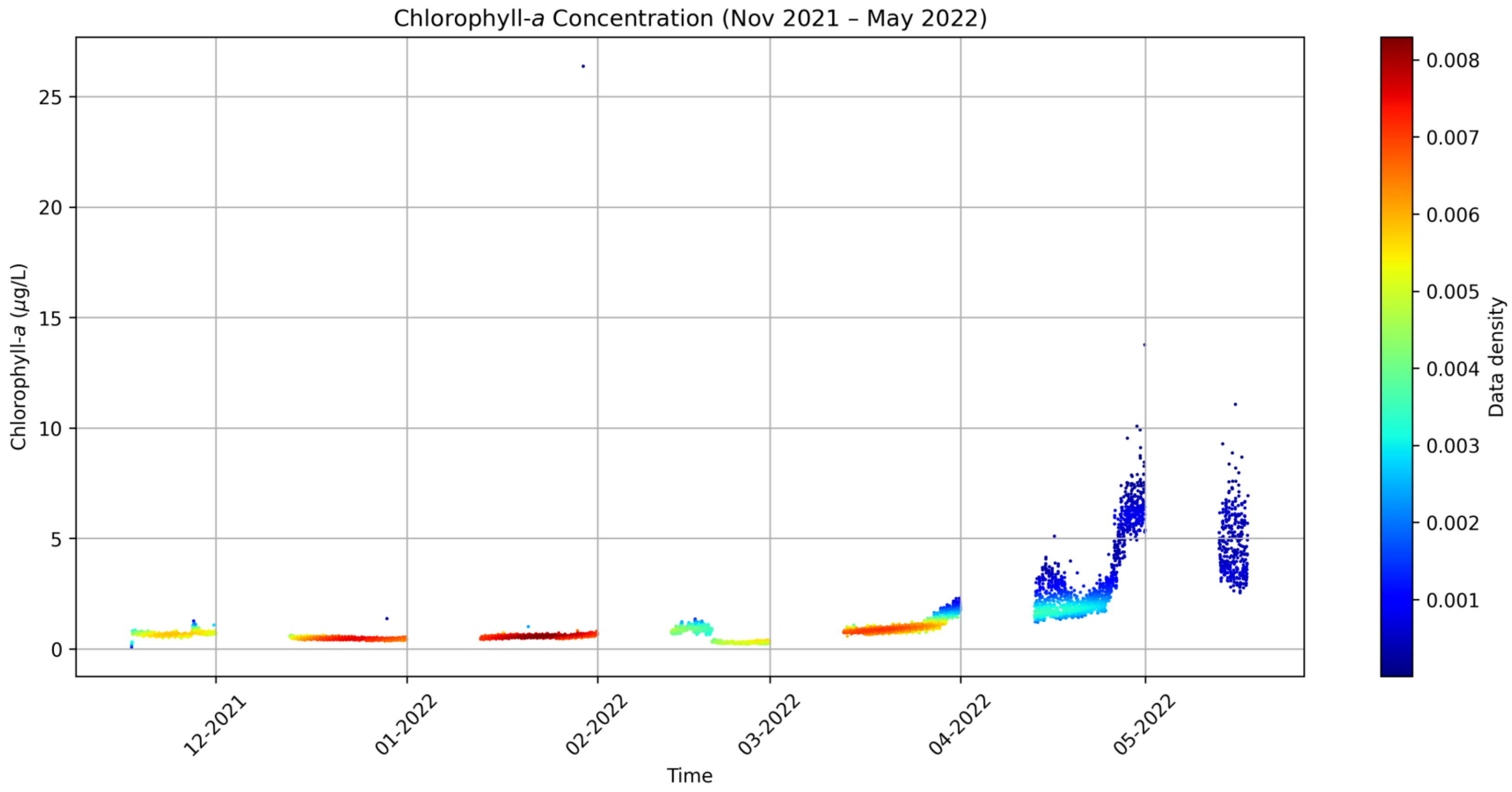


Fig. 2 Time series of chlorophyll-a concentration (μg/L) from November 2021 to May 2022. A colourbar from blue (lower) to red (higher) is used to represent the local data density.

Sentinel-3 Data

Chl-a products

Preprocessing: Apply WQSF masks (cloud, glint, invalid)

- C2RCC (L1)
- OC4Me (L2)
- Neural Network (NN) (L2)

Comparison

- OC4Me: higher spatial variability
- NN: smoother, less noise

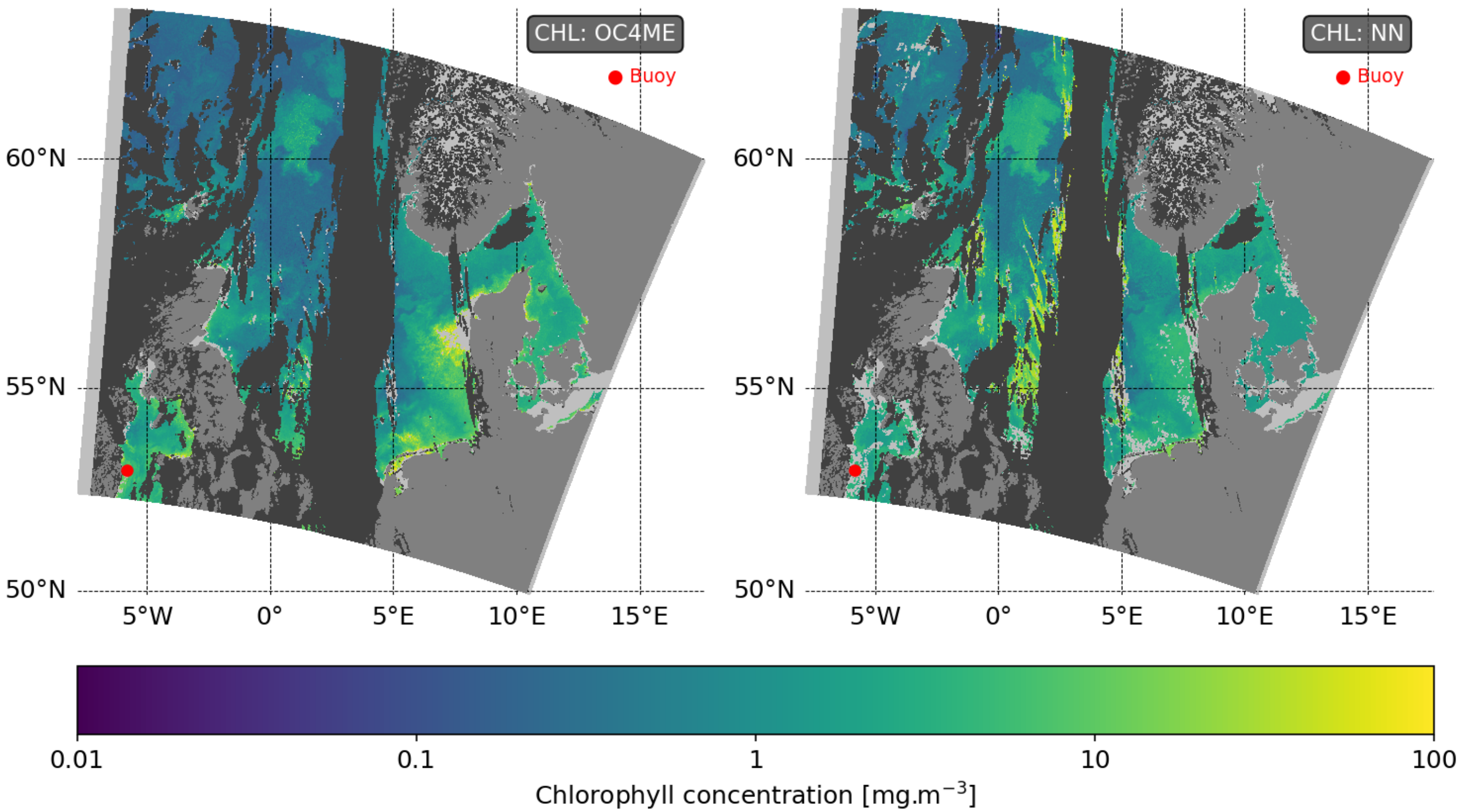
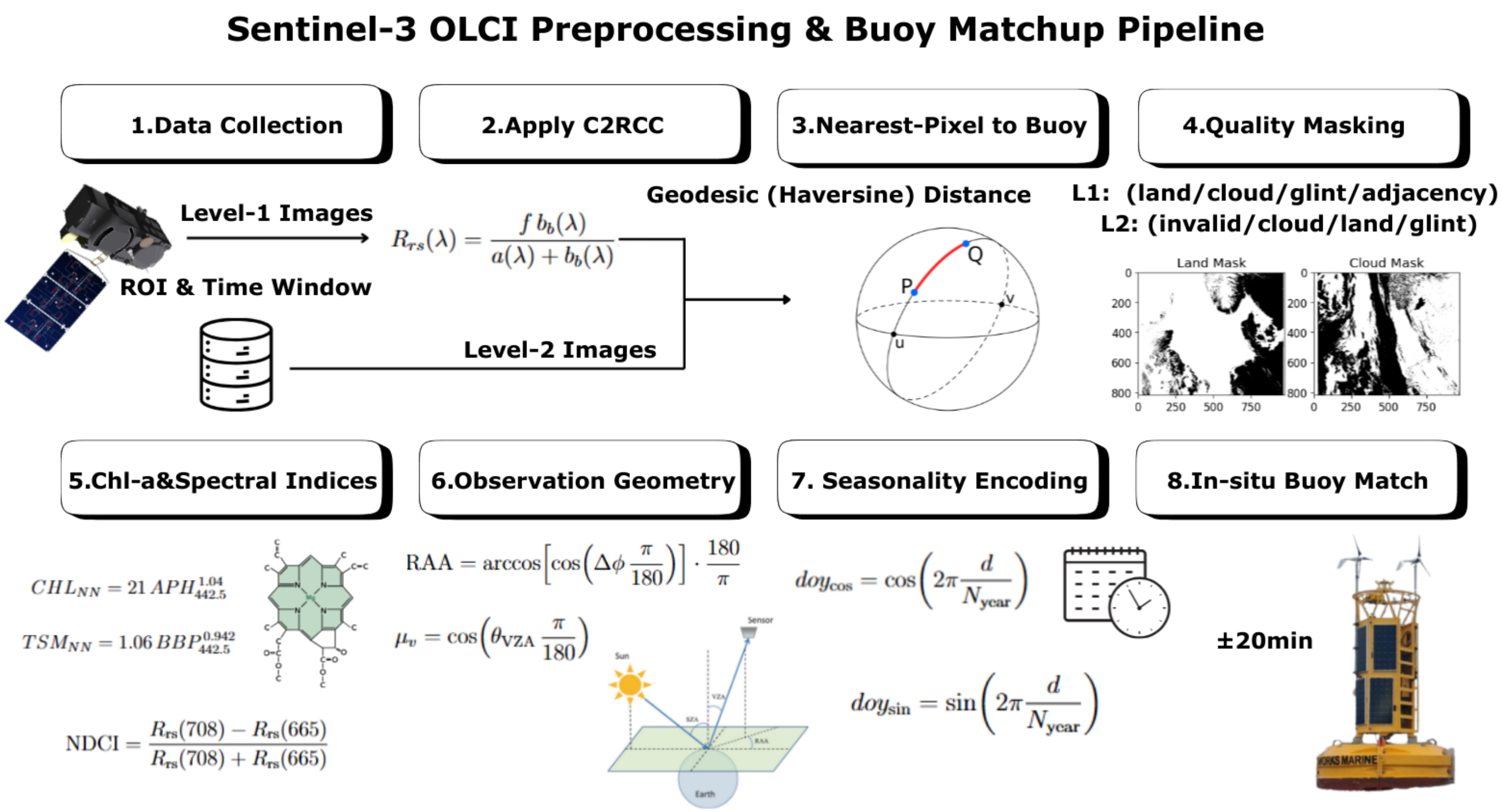


Fig. 3 Comparison of OLCI chlorophyll-a concentration maps at full resolution (19/04/2022). The chlorophyll-a concentration is shown on a $\log_{10}$ scale. Code for the plot adapted from EUMETSAT.

Methodology



Experiments

Models

- Random Forest
- XGBoost
- Support Vector Regression

suitable for small datasets

Method:

- Grid search (hyperparameter tuning)
- k-fold cross-validation

Input features for C2RCC

- Chl-a
- Total Suspended Matter concentration
- Reflectance (443–555 nm)
- Normalized Difference Chlorophyll Index (NDCI)
- Seasonal encoding (DOY sin/cos)

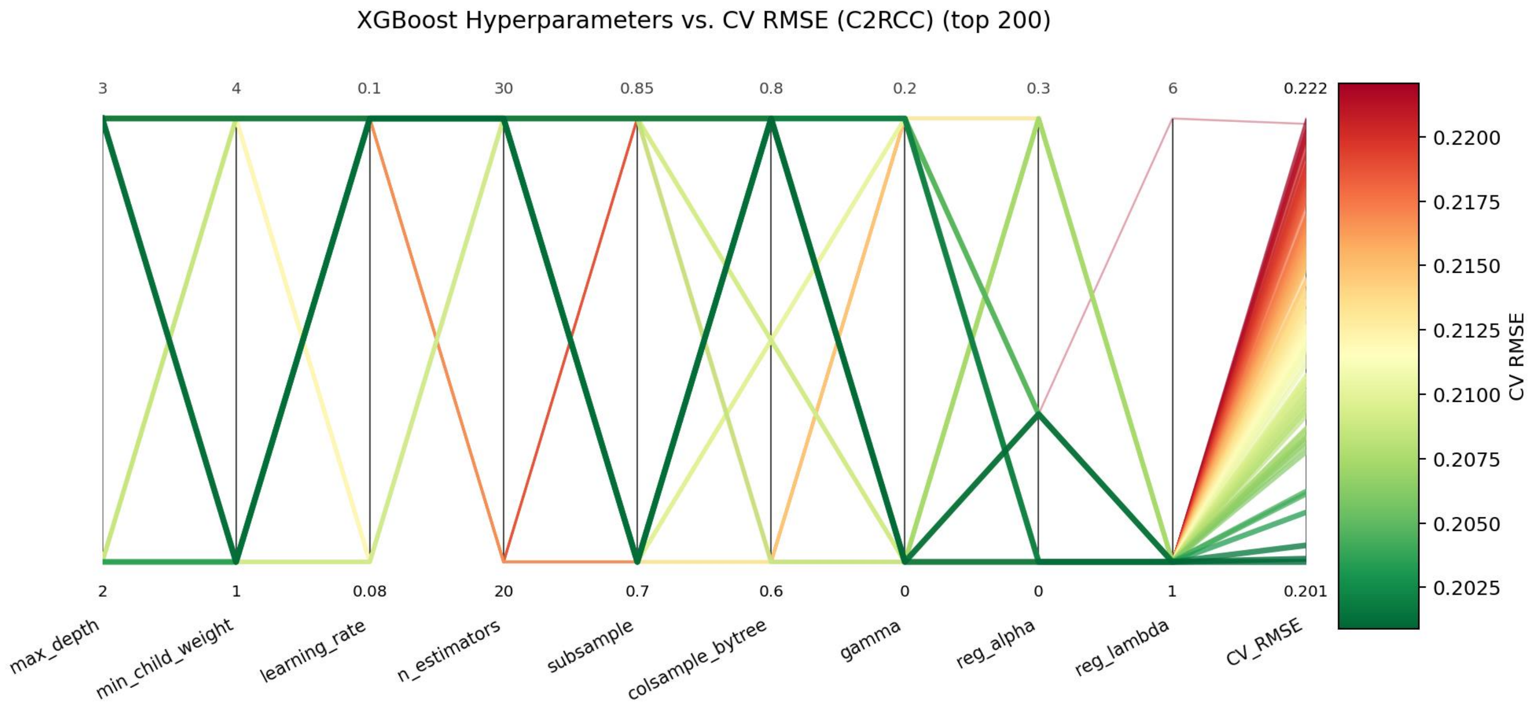


Fig. 4 Parallel coordinates plot of the top 200 XGBoost hyperparameter configurations for the C2RCC dataset, ranked by cross-validation RMSE. Each line corresponds to one configuration, colouredfrom green (low RMSE) to red (high RMSE).

Results

Results

- Compared: XGBoost, RF, SVR
- XGBoost best performance

Metrics (Test) - C2RCC XGBoost

- $R^2 = 0.93$
- RMSE = 0.289 mg m^{-3}
- MAE = 0.162 mg m^{-3}

Observation

- comparable train-test metrics
- small residuals
- indicate no evident overfitting

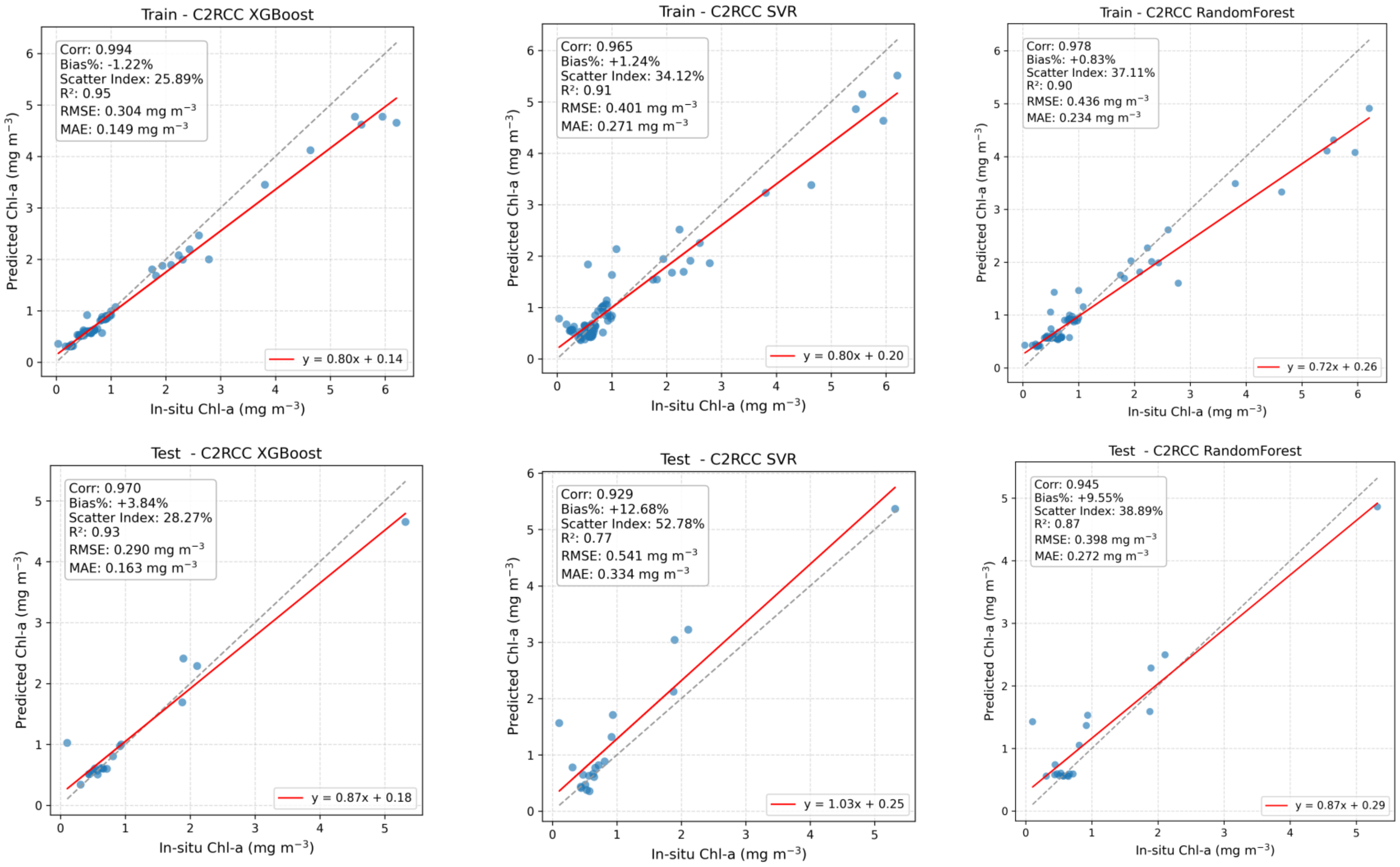


Fig. 5 Model performance for the C2RCC dataset: training and test results

Model Performance

Model performance

- all models performed well
- XGBoost outperformed the other two models
- strong generalisation (train similar to test)
- SVR outperformed RF

C2RCC

Model	Split	Evaluation Method		
		R^2	RMSE	MAE
XGBoost	Train	0.949	0.303	0.149
	Test	**0.932**	**0.289**	**0.162**
SVR	Train	0.911	0.400	0.271
	Test	0.765	0.540	0.334
Random Forest	Train	0.895	0.435	0.234
	Test	0.872	0.398	0.271

OC4Me

Model	Split	Evaluation Method		
		R^2	RMSE	MAE
XGBoost	Train	0.938	0.334	0.167
	Test	0.808	0.488	0.299
SVR	Train	0.895	0.436	0.218
	Test	0.870	0.402	0.264
Random Forest	Train	0.903	0.419	0.212
	Test	0.786	0.516	0.285

NN

Model	Split	Evaluation Method		
		R^2	RMSE	MAE
XGBoost	Train	0.955	0.284	0.138
	Test	0.878	0.388	0.232
SVR	Train	0.923	0.372	0.213
	Test	0.811	0.484	0.338
Random Forest	Train	0.894	0.437	0.229
	Test	0.847	0.435	0.284

Table 1. Comparison of models for the C2RCC dataset (train and test) using R^2, RMSE and MAE.

Explainable AI

Explainable AI (SHAP)

- Identifies feature importance and direction

Top contributors

- Seasonality (DOY sin/cos)
- Observation geometry

Secondary contributors

- Reflectance, NDCI

Strong seasonal cycle of chlorophyll-a values

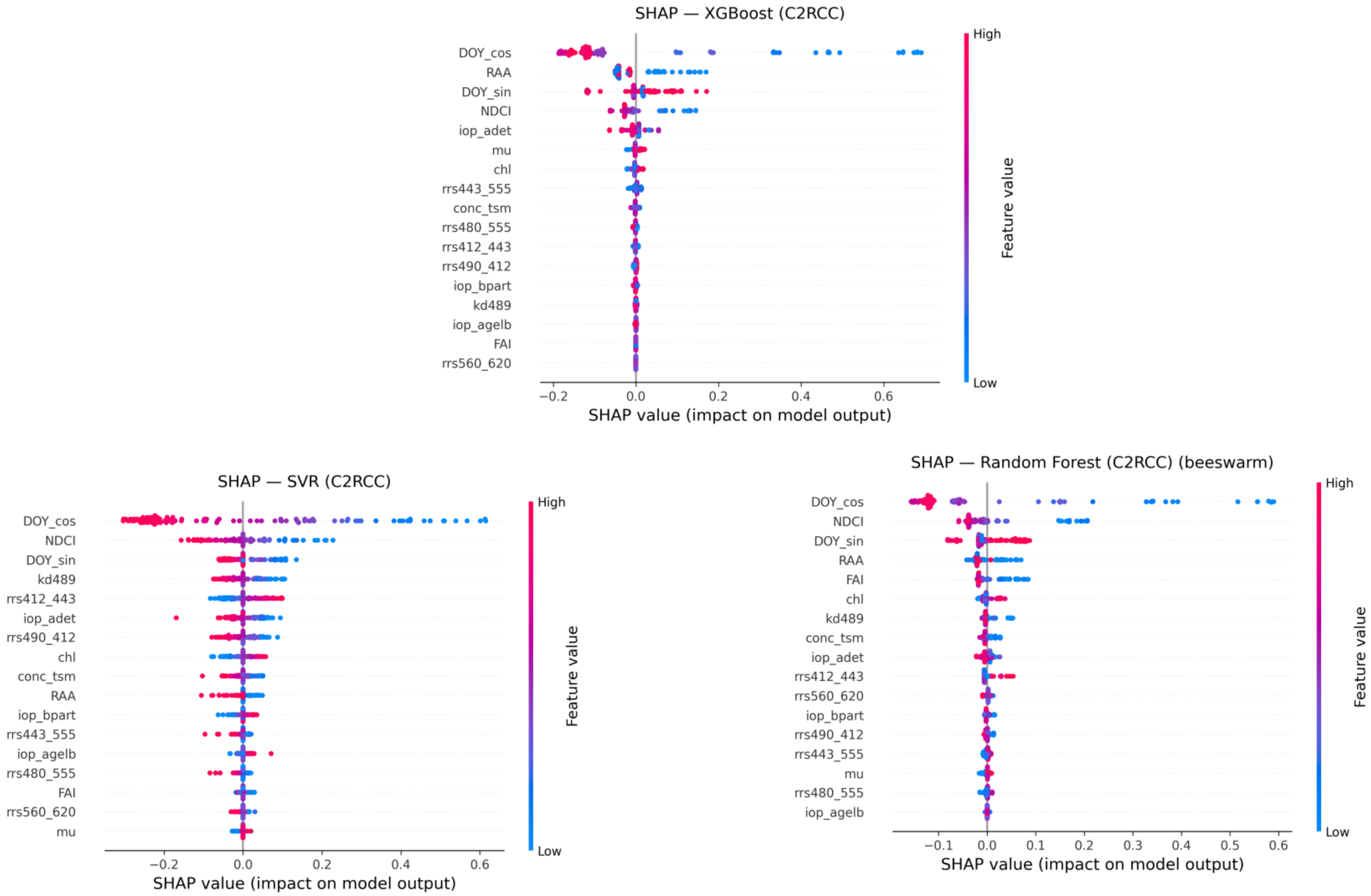


Fig. 6 SHAP plot for the C2RCC dataset with XGBoost, SVR, RF models

Conclusion

1. Best performing setup: C2RCC with XGBoost. Explained $R^2 = 0.93$ of the variance with low prediction error

2. Top contributors: Seasonality, Geometry viewing. Phytoplankton follows a seasonal pattern

3. Generalisation: Comparable train-test performance and generally small residuals suggest no overfitting

4. Error pattern: No systematic bias, only that higher concentrations are mildly underpredicted

Future Work

1. Enrich model input e.g. bathymetry and mixed layer depth
2. Deploy a grid of buoys to have multiple validation points
3. Implement Bayesian modelling for uncentainty
4. Improve seasonal coverage (year-round observations)

TECHWORKS
MARINE

Thank you for listening