# Unified Access to DEIMS In-Situ Data via STAC for Scalable Earth Observation Machine Learning

Ignacio Masari, Adrian di Paolo, Christoph Reimer

*EODC Earth Observation Data Centre for Water Resource Monitoring GmbH*

**STAC discoverability** | **Per-measurement spatio-temporal indexing** | ML-ready data access workflow | Deep learning compatible pipeline | Scalable compute via EODC's clusters

## The Ground-Truth Bottleneck in EO ML

Large-scale machine learning for Earth Observation depends on harmonised, discoverable labelled data. While satellite archives are well structured, in-situ environmental measurements remain fragmented and difficult to integrate into ML workflows.

Within Destination Earth (DestinE), scalable model development requires machine-readable, interoperable ground-truth data.

**Current limitations:**

- Heterogeneous metadata
- Limited spatial/temporal queryability

This creates a **bottleneck** for reproducible ML pipelines.

## STAC Integration of DEIMS-SD

We expose 500+ monitoring sites from DEIMS-SD through a fully compliant STAC interface.

**Features**:

- Catalog → Collection → Item structure
- Either per-measurement or per time-series temporal indexing
- Native spatial geometries
- Public API endpoint
- Metadata preserved in original units

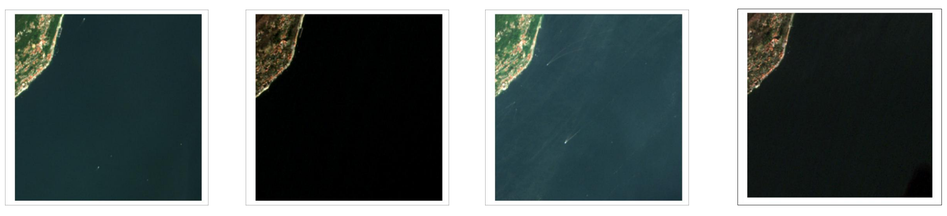
**Users can query existing ground-truth before planning new campaigns.**

## From STAC Query to ML: Lake Maggiore Example

We demonstrate an end-to-end workflow using Chlorophyll-a (Chl-a) in-situ measurements (2015–2018) paired with Copernicus Sentinel-2 L2A imagery over Lake Maggiore, Italy.

**Setup**

- Monthly measurements (~40 samples)
- 256×256 Sentinel-2 patches centred at station
- Regression to single Chl-a value
- Temporal Leave-One-Year-Out validation



| Date | 06/08/2015 | 16/01/2016 | 11/06/2017 | 11/12/2018 |
|---|---|---|---|---|
| True Chl-a (µg/L) | 2.9 | 0.8 | 4.93 | 1.65 |
| Predicted Chl-a (µg/L) | 3.04 | 0.87 | 4.24 | 1.48 |

## Foundation Model Integration

We employ Clay (v1.5) to illustrate compatibility with modern ML pipelines.

**Approach**:

- Frozen backbone
- Patch embeddings extracted
- Lightweight MLP regression head
- No fine-tuning

This highlights rapid integration of foundation models within a STAC-driven data workflow.

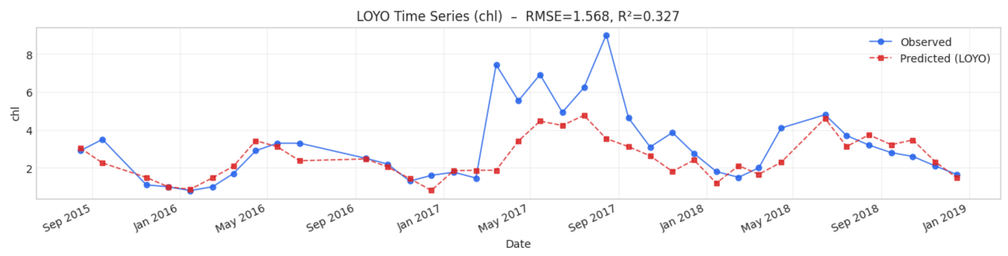
## Scalable Compute via EODC

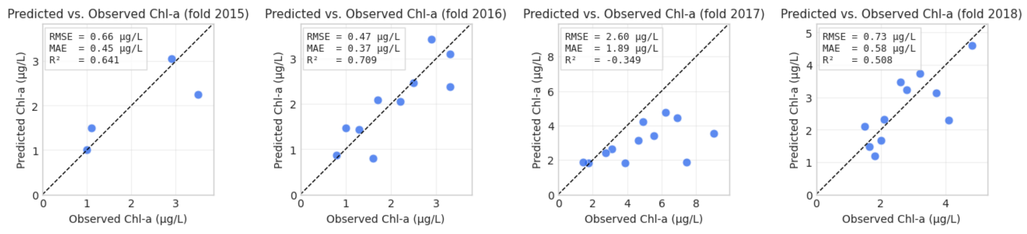
Experiments were executed on GPU infrastructure at EODC.

- Single NVIDIA H200 MIG (1/7) slice
- Production-ready STAC deployment
- Scalable GPU resources
- On-demand access framework under development

Researchers without in-house HPC can discover data, assemble training sets, and run large deep learning model pipelines within a unified ecosystem aligned with DestinE.

## Performance on the case study




Repository

DEIMS-SDR

EODC