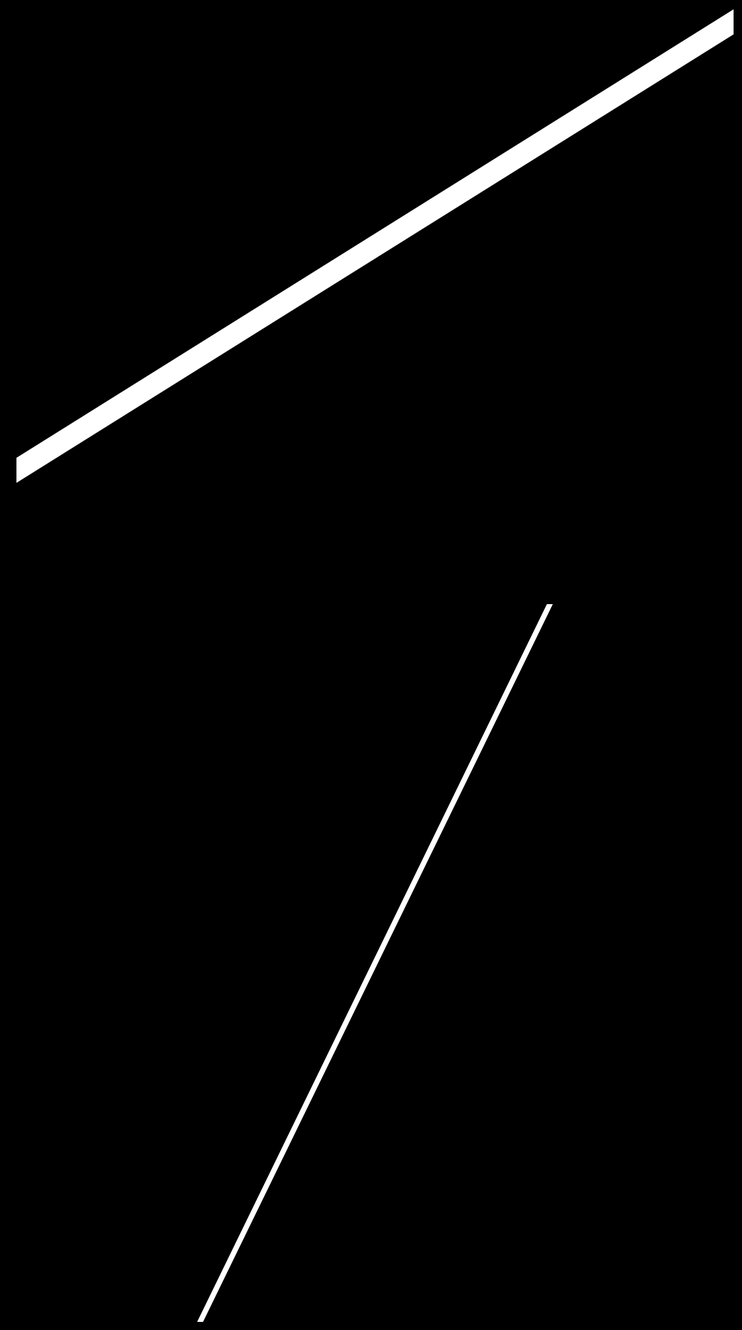
GAIN SUMMER SCHOOL

Data Driven Modelling for Aquaculture

Fearghal O'Donncha Ph.D
feardonn@ie.ibm.com

This project receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 773330 (GAIN)

IBM Research: Inventing What's Next



3000
Researchers

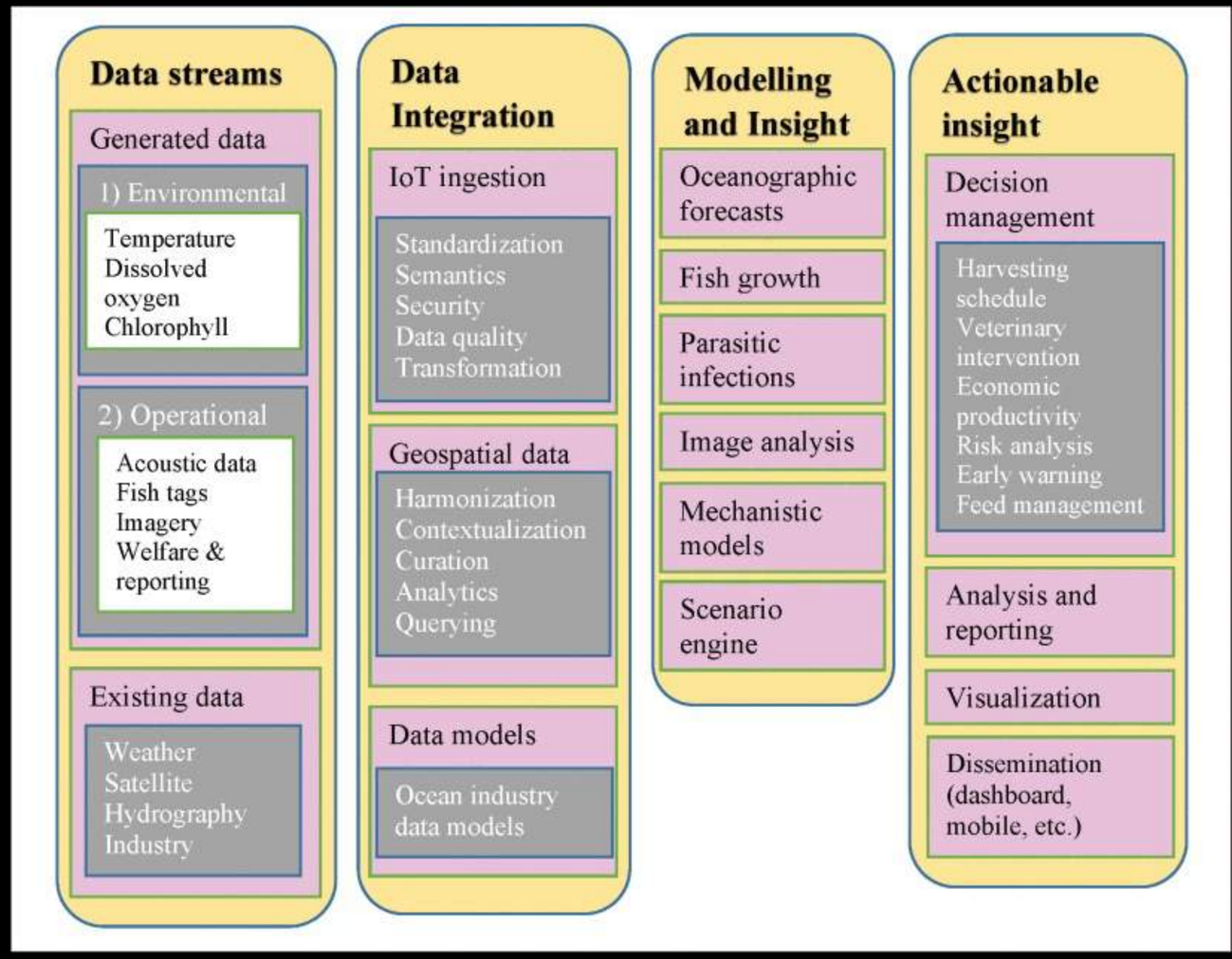
19
Locations

6
Continents

6 Nobel Laureates
10 Medals of Technology
5 National Medals of Science
6 Turing Awards

Monitor, Analyse, Interpret, and Decide

Precision Aquaculture



Forecasting environmental conditions - Aquaculture

Copernicus Marine Service
1) a unique service

marine.copernicus.eu

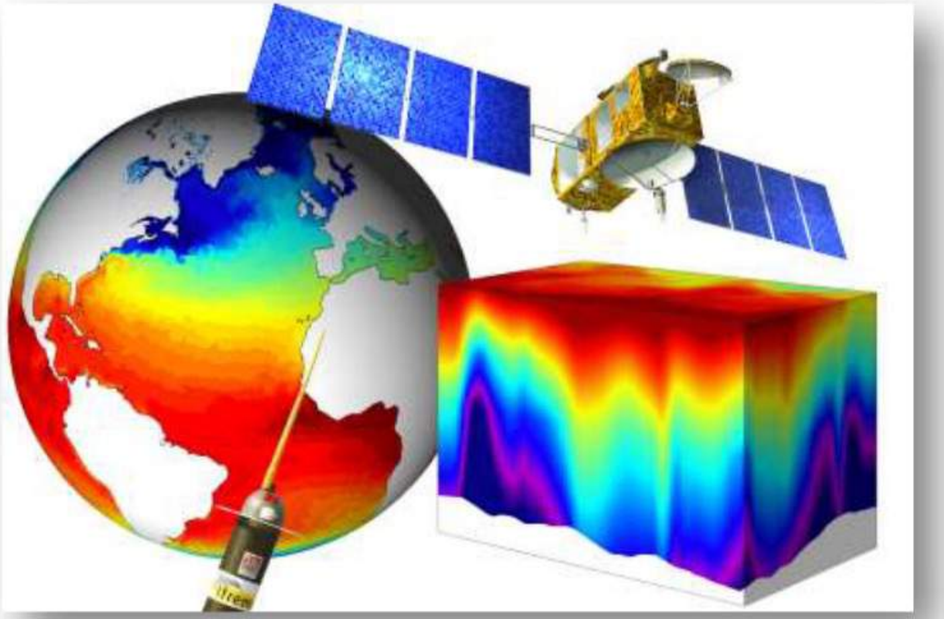
EASY ACCESS TO OPERATIONAL OCEANOGRAPHY PRODUCTS FOR ANYONE

ONE-STOP-SHOP-WINDOW

WORLD OCEAN / EUROPEAN SEAS

ESSENTIAL OCEAN VARIABLES

Copernicus Marine Service
2) an integrated information



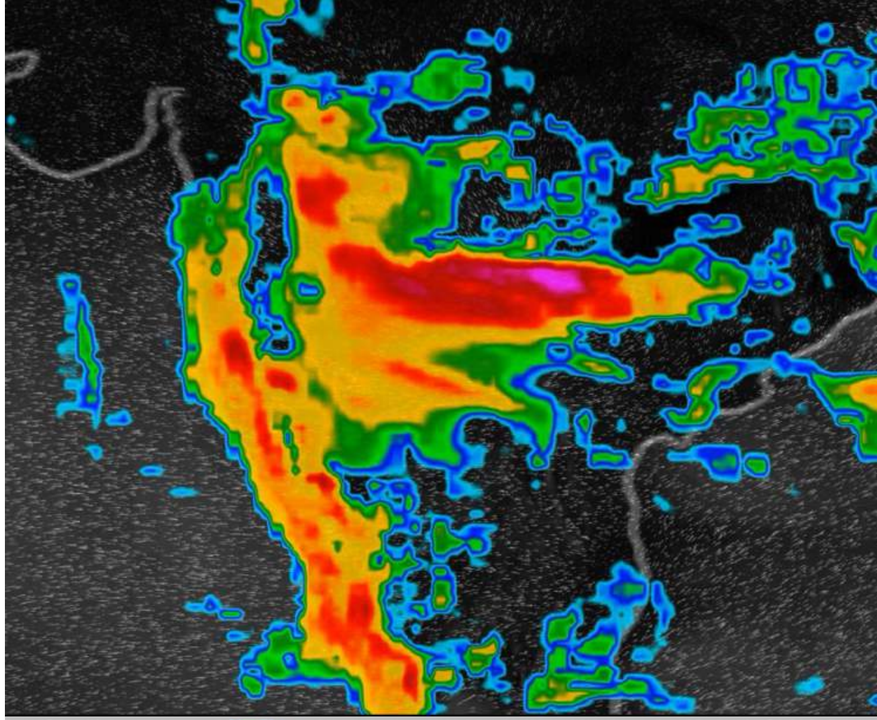
OPERATIONAL OCEANOGRAPHY BUILT UPON MULTIPLE SOURCES OF INFORMATION

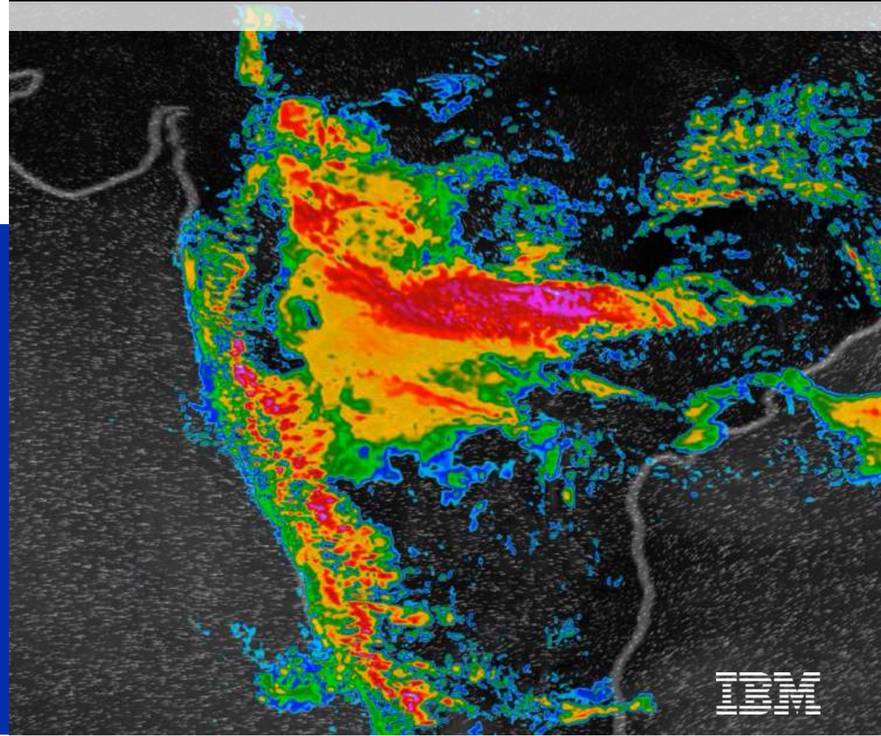
OBSERVATIONS AND MODELS
PHYSICS AND BIOGEOCHEMISTRY
REAL-TIME AND REANALYSES

The Weather Company
An IBM Business

Global Hi-Resolution Weather Forecasting (GRAF) System 2

- Most models are at 13-15km resolution and updated just 2-4 times daily
- GRAF will run at 3km for populated areas (~30% of the globe) and at 4km for ConUS and western Europe
- 5 minute time-steps vs hourly
- First-ever operational global weather model to run on GPU-accelerated servers
- For short term models, expected to be the BEST





GRAF brings the resolution once limited to the US, Japan, and Western Europe to outstanding levels

Geospatial Analytics

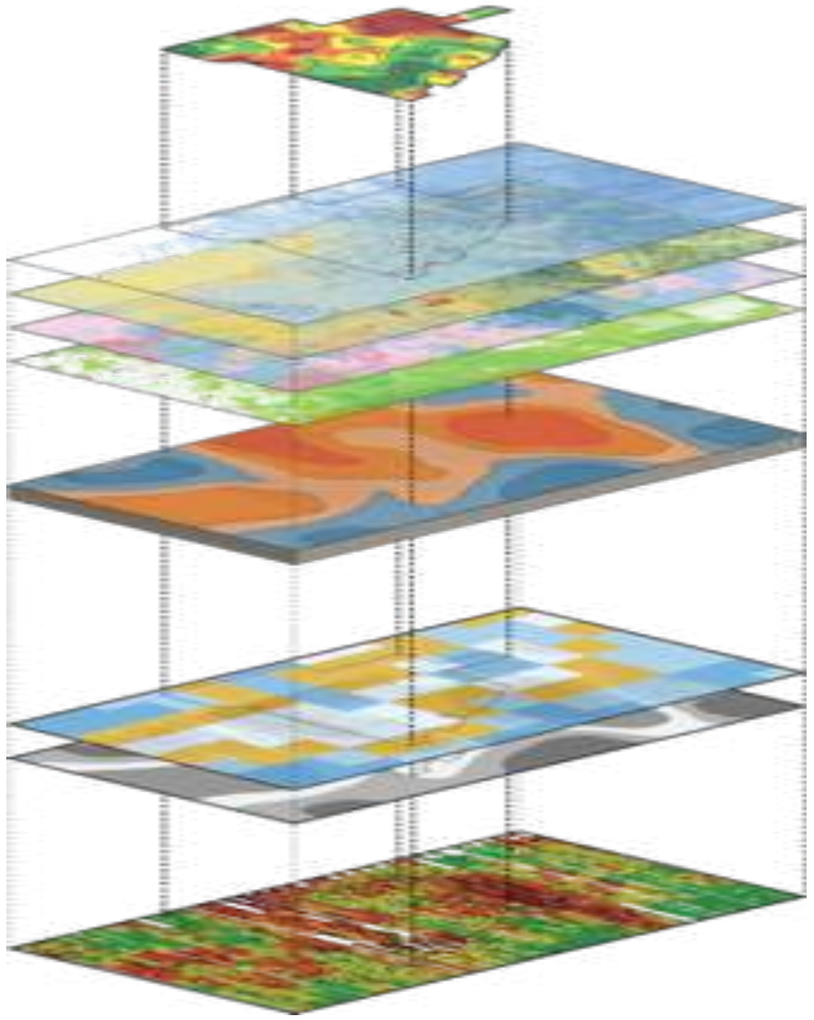
Aggregates and analyzes multiple layers of geospatial data and lets users query across the combined layers without loss of resolution. Data and output from PAIRS supports the training of new machine learning models.

Retrieve relevant information with flexible geospatial and temporal resolution

Create new layers that provide critical insight from queries/analytics which can correlate data across public & private data

Select from library of existing layers (e.g. satellite) to combine with your data

Ingest your own data, manage, normalize and align to common grid for location & time



https://www.ibm.com/products/environmental-intelligence-suite/geospatial-analytics

Aquaculture industry requirements analysis

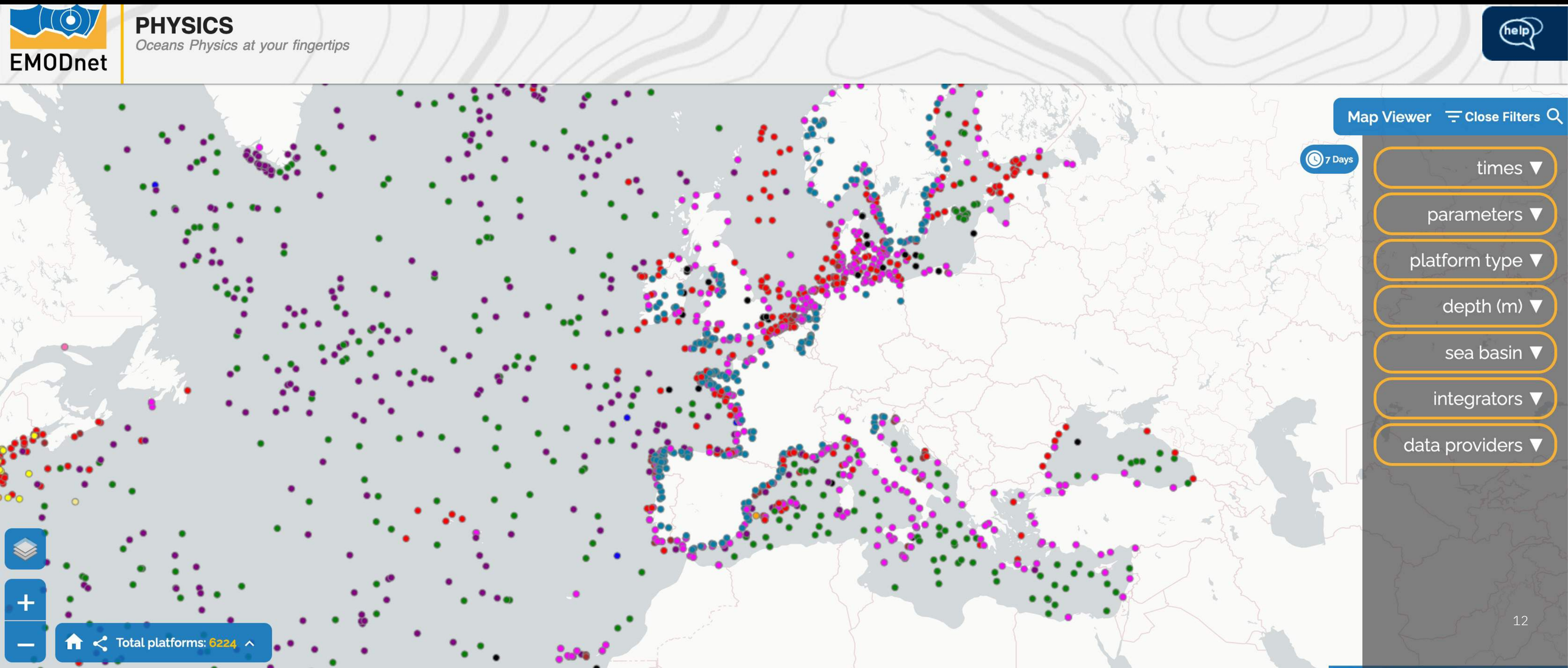
Copernicus Training Workshop Report

“The monitoring of oxygen levels, pollution drift forecast, algal bloom monitoring and eutrophication was regarded of highest relevance (56% of respondents gave this the highest score), followed by the use of data for the management of environmental risks (52 % of respondents gave this a 5-score)”.

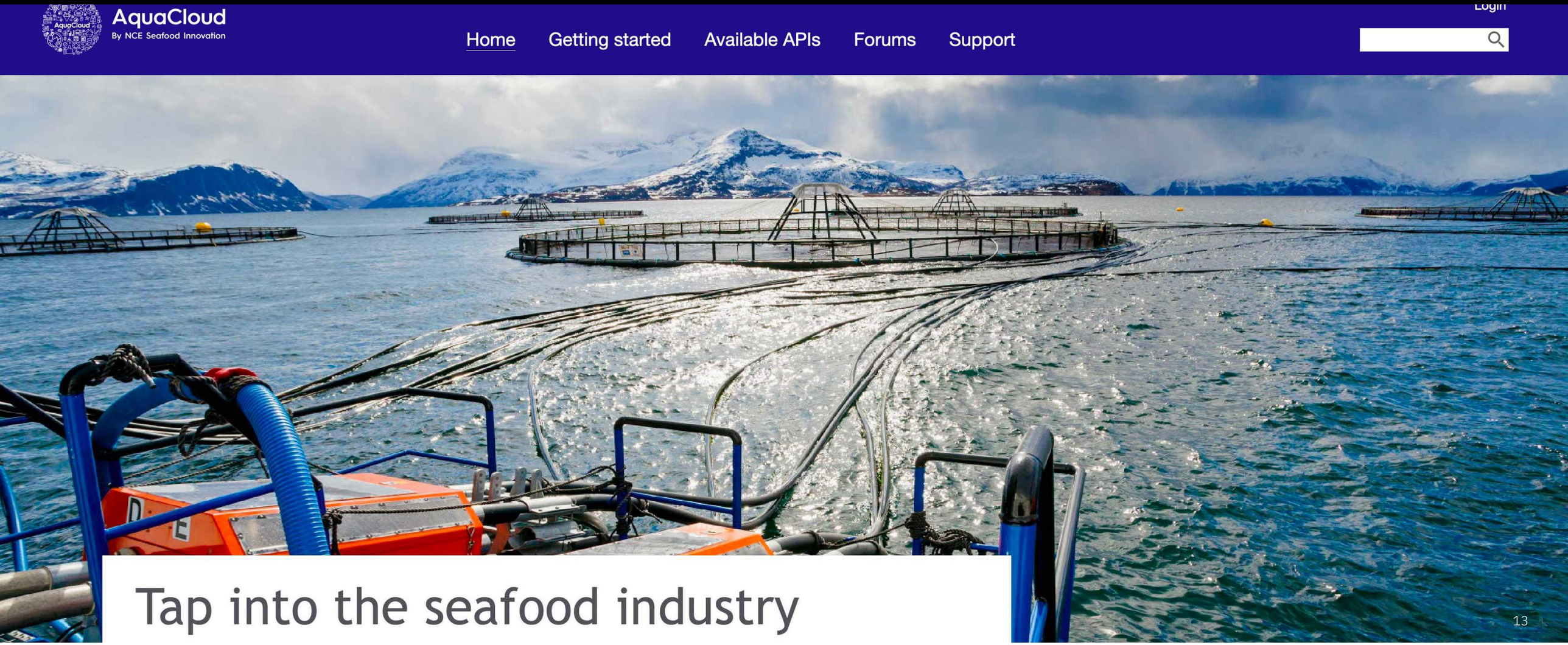
“Attendants from the aquaculture sector required high resolution coastal data sets (100 m – 1 m) and a short delivery time”.

https://marine.copernicus.eu/sites/default/files/wp-content/uploads/2019/11/Workshop-CMEMS4AQUACULTURE-Athens-SEP2019.pdf

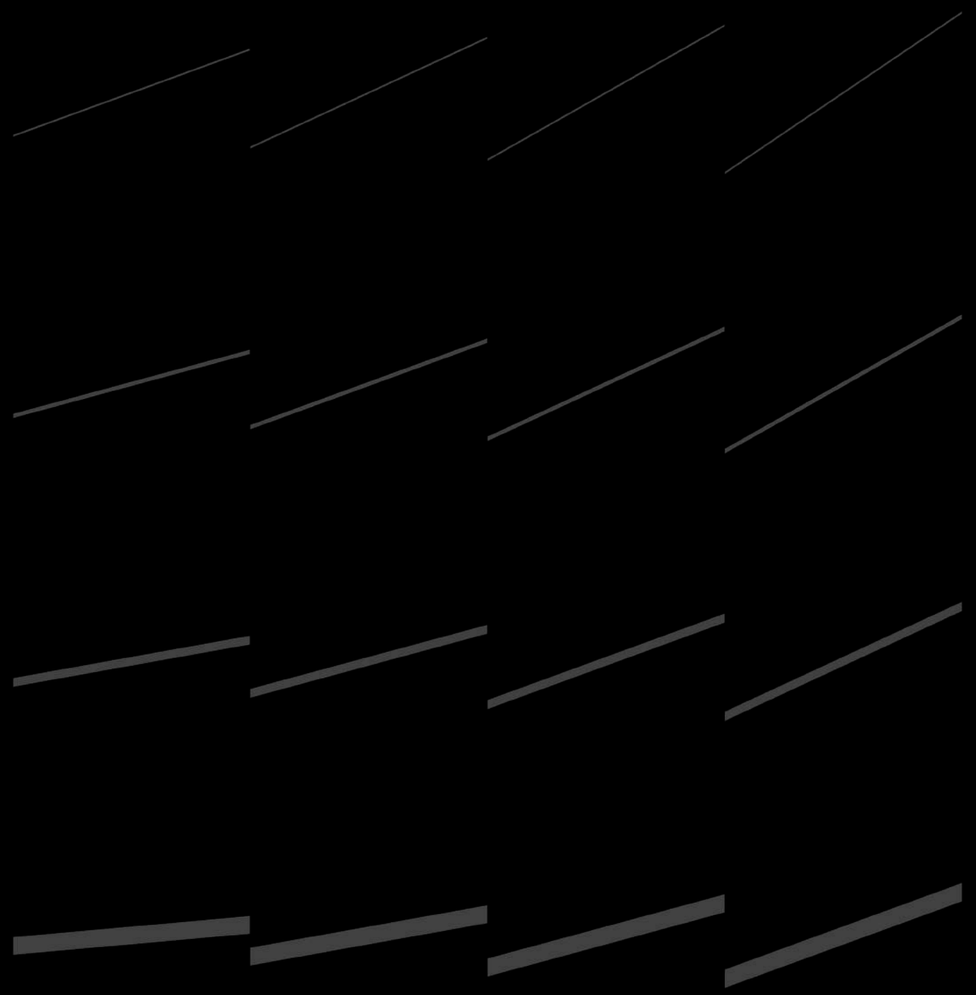
Ocean Observation data



Datasets specific to aquaculture



Machine Learning forecasting of conditions at requisite scale



Pilot Deployments: The geographical sites

Species

Canada

Norway

Scotland

Ireland

Poland

Italy

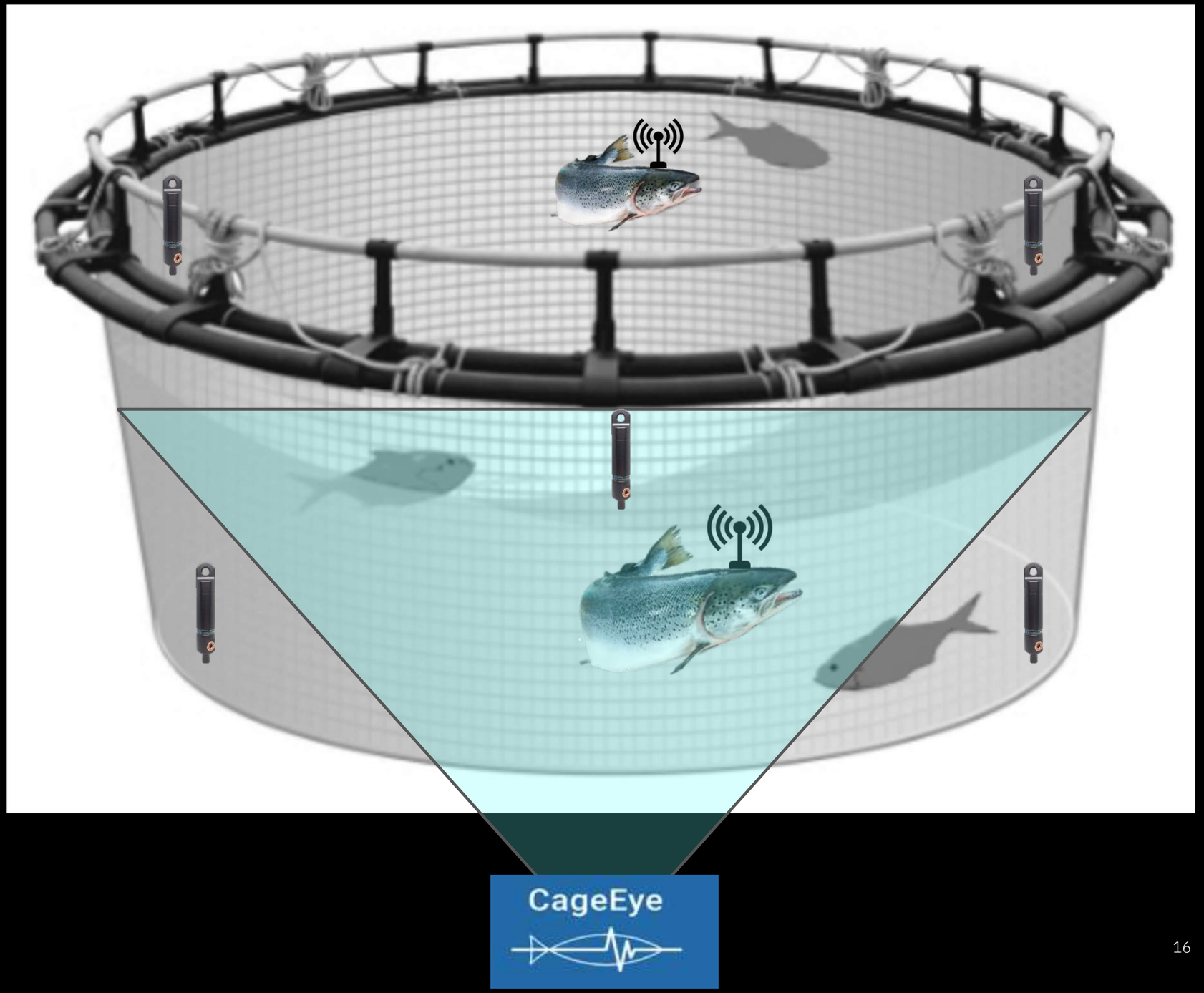
Portugal

Spain

China

Aquaculture 4.0

Multiple sensors sampling realtime environmental conditions at different cage positions

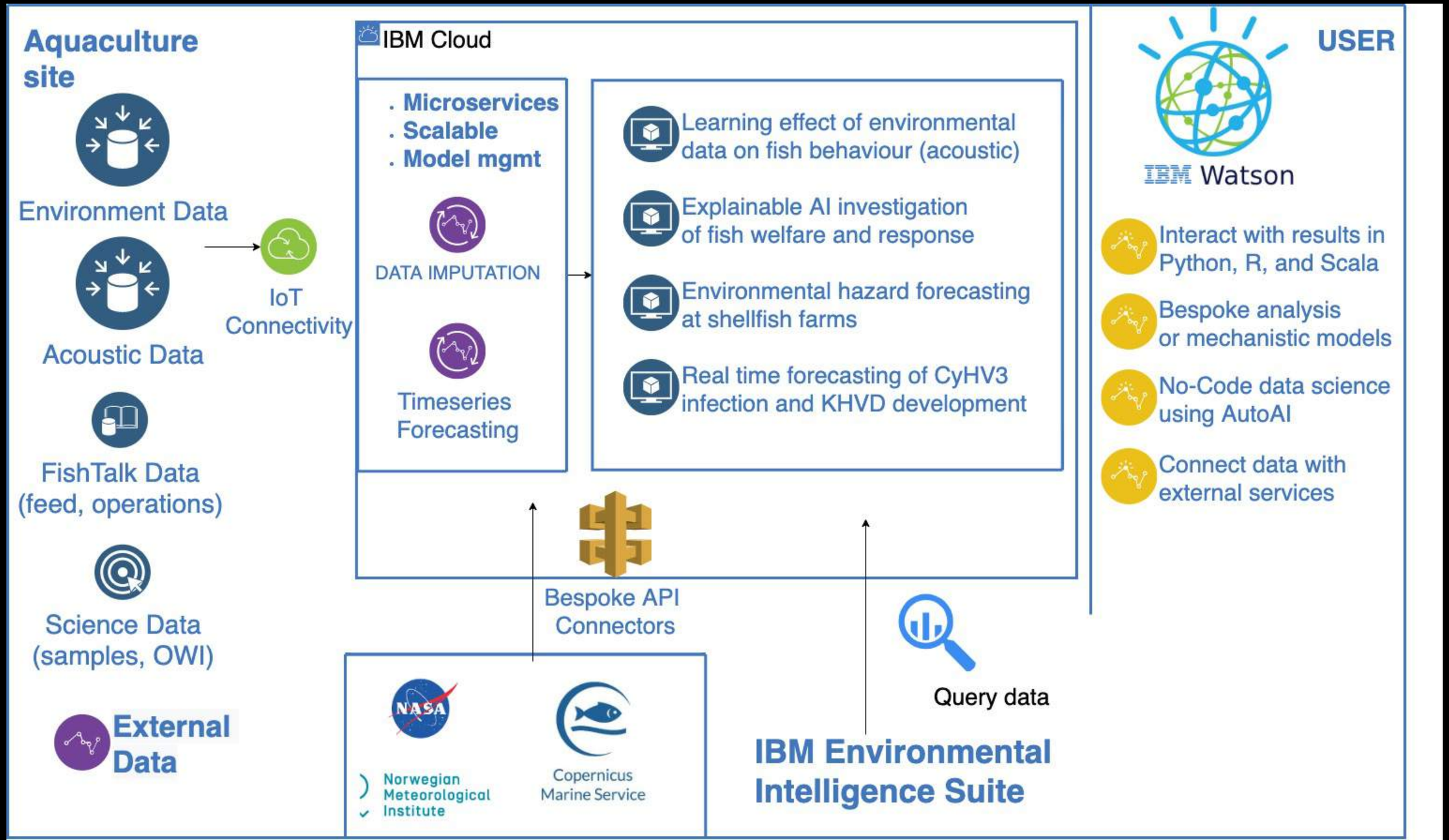


Data-informed decision for aquaculture

Four overlapping components fundamental to addressing aquaculture pain points:

1) integrate sensor data from each pilot partner site to a secure cloud service;
2) augment with curated data from pertinent external sources (satellite, ocean model, weather, etc.) into an accessible, easily queried, format;
3) develop a suite of data-driven and mechanistic models to provide predictions of pertinent conditions and extract insight from data; and
4) disseminate to the end-user in a secure manner that interfaces with other operations.

Precision Aquaculture Components



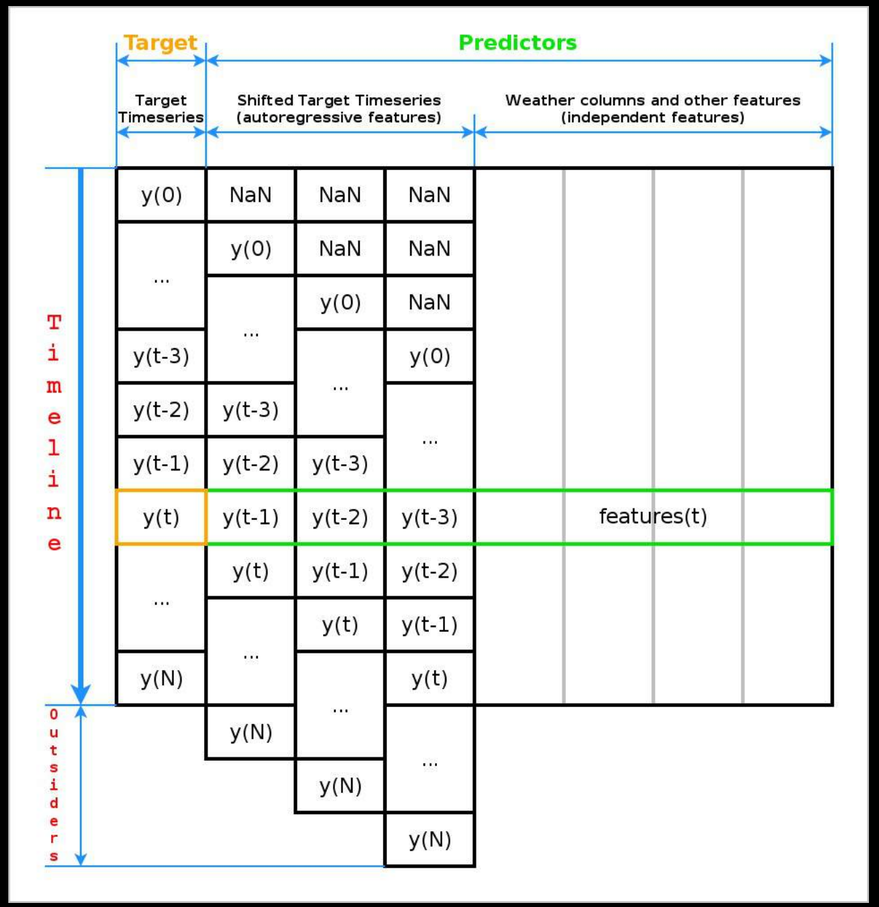
Outlier and anomaly detection

As number of sensors increase, more automation is critical

- While current pipelines expect pre-loaded data into set of uniform matrix arrays, IoT data requires a pipeline that can pull data from multiple sources, at different temporal positions and resolutions on-the-fly.
- Requires a series of transformations to make consumable for model:
  - Interpolation, rescaling, imputation of missing data
  - PCP and LR based approaches for sensor drift and outlier detection

Feature engineering and hyperparameter optimisation

- Robust feature engineering of autoregressive data (subject to data gaps, uncertainty, etc.)
- Integration of data from multiple external sources (TWC, PAIRS), and models (other ML models)
- Scalable feature testing and hyperparameter optimisation using pywren.
- Pywren is an open source project whose goals are massively scaling the execution of Python code and its dependencies on serverless computing platforms and monitoring the results.



Machine learning forecasting models

- Implemented number of time series forecasting models from statistical (GAM, ARIMA), machine learning (Random Forest, XgBoost), and deep learning (NN)

- Easily prescribe different model inputs (the data inputs of salmon farm in Norway is different to a carp farm in Poland)
- Wrap within the hyperparameter optimisation routine to allow automation of model pipeline
- Weighted ensemble aggregation approaches based on observation[1]

O'Donncha, Fearghal, et al. "Ensemble model aggregation using a computationally lightweight machine-learning model to forecast ocean waves." *Journal of Marine Systems* 199 (2019): 103206.

Increased complexity makes model management critical

- Multiple variables (temperature, dissolved oxygen, Chl-a, ...) from multiple locations (different farms, cages, location within cage).
- Different data cleaning and imputation approaches
- Multiple algorithms to chose from (GAM, ARIMA, Random Forest, Recurrent Neural Networks, etc.).
- Different temporal (seasonal) and model drift performance statistics

Scalable, automated (robust to disparate, noisy, and gappy data) timeseries forecasting models that can be readily deployed to multiple (tens-to-hundreds) sensor datasets

Chen, Bei, et al. "Castor: Contextual IoT time series data and model management at scale." *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2018.

Requirements summary

1. an efficient pipeline for ingesting IoT time series data in real time
2. a scalable, hybrid data management service for both time series and contextual data
3. a versatile semantic model for contextual information which can be easily adapted to different application domains
4. an abstract framework for developing and storing predictive models in R/Python (or the particularly exciting Julia language)
5. deployment services which automatically train and/or score predictive models upon user-defined conditions.

Simplify the life of the domain scientist

Chen, Bei, et al. "Castor: Contextual IoT time series data and model management at scale." *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2018.

Time Series Forecasting

Model Comparison

Transform input to functional time-series data

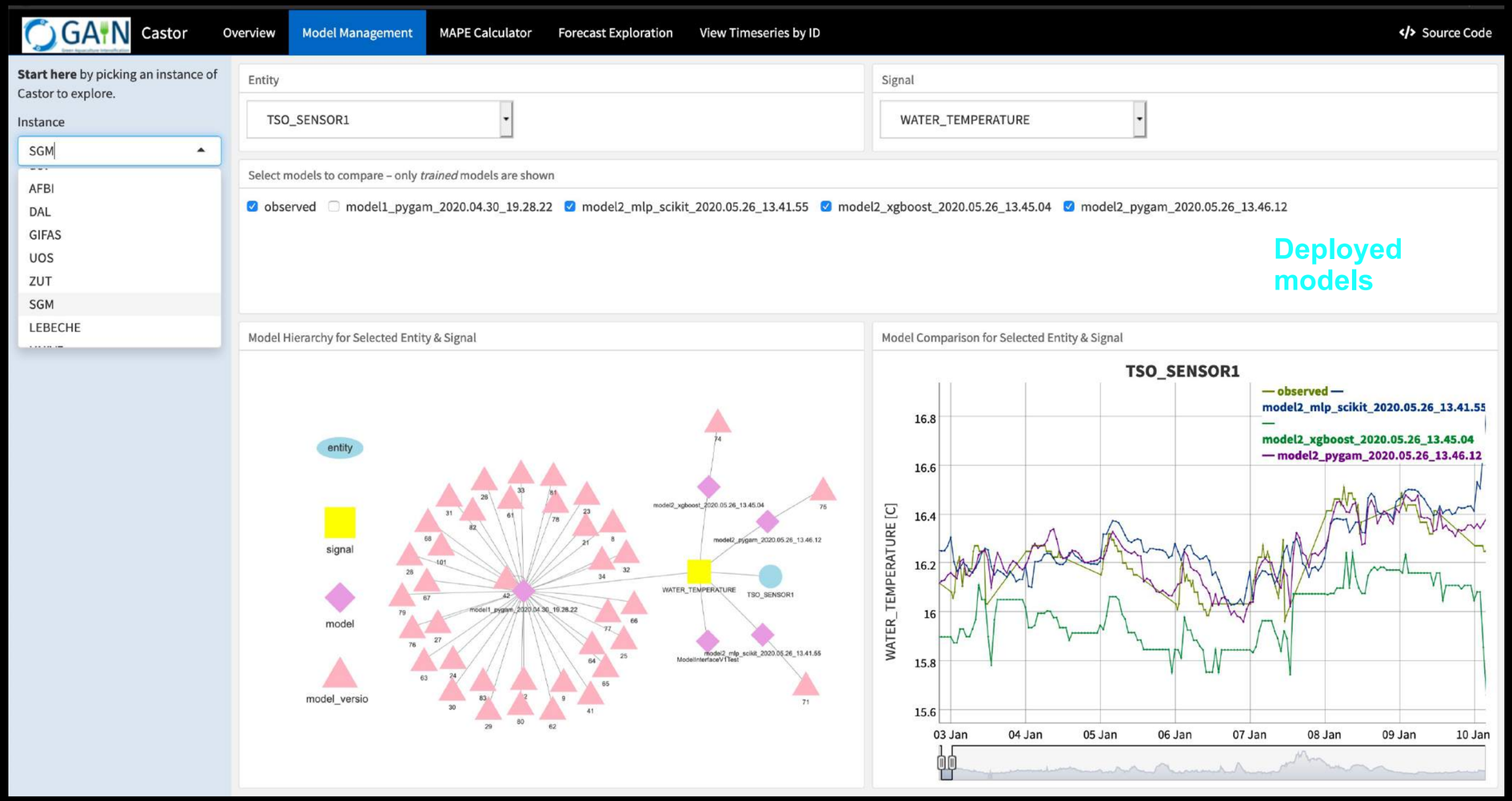
Product classification by time-series signature

Curated external data signals

“Top 10” external signals impacting each product

Enhanced LSTM using autoencoder and FFT techniques (uncover hidden patterns)

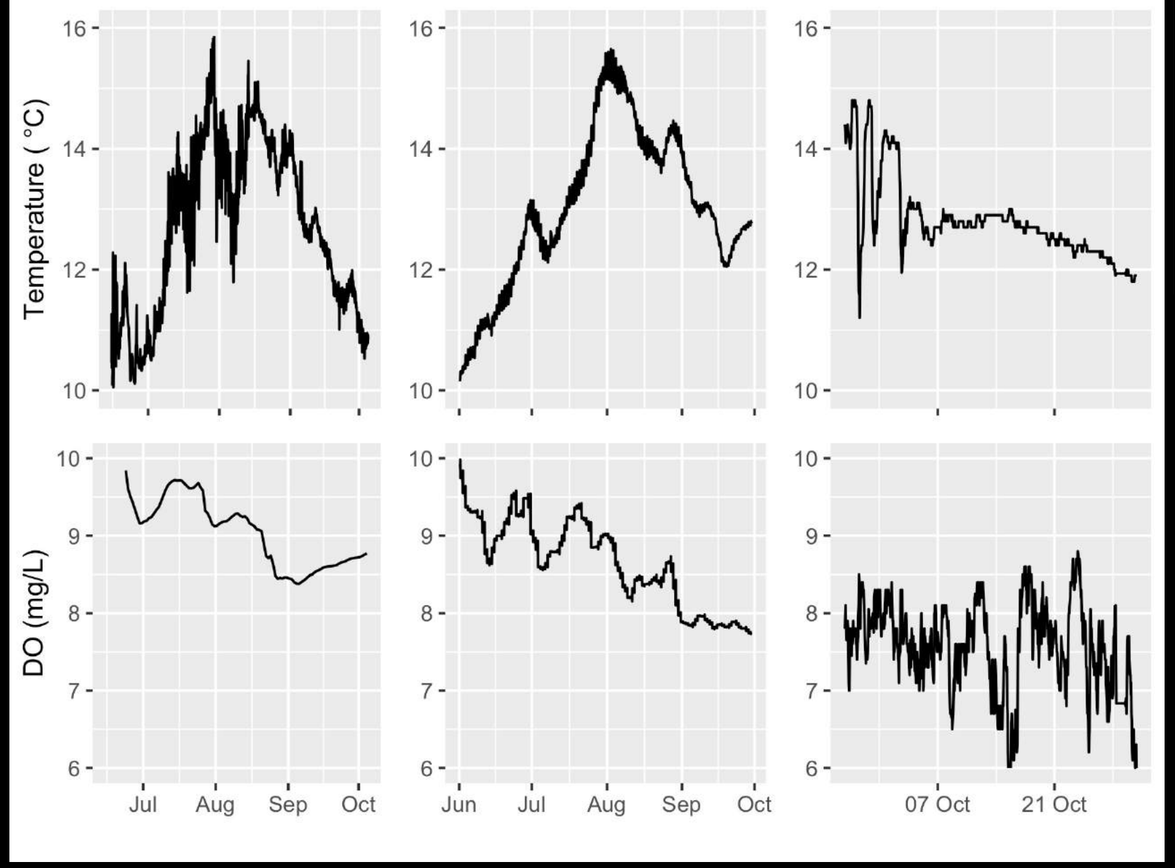
Used coarse-grained ensemble to select best model for each observation



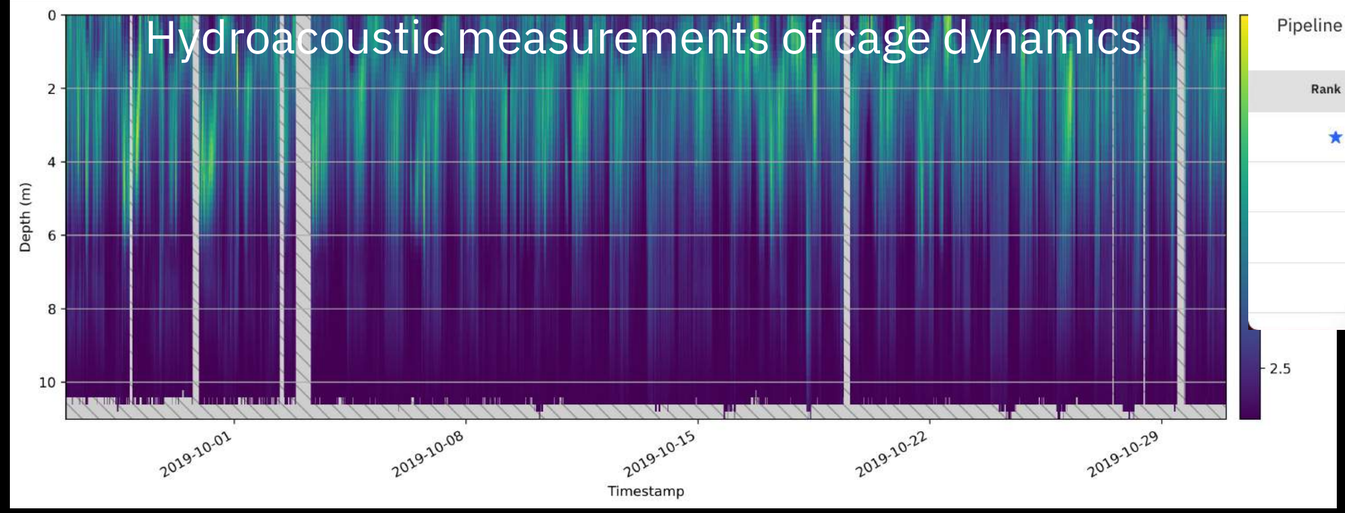
Knowledge on the environment informs management

Data Driven Insight Into Fish Behaviour and Their Use for Precision Aquaculture

Fearghal O'Donncha[1*], Caitlin L. Stockwell[2], Sonia Rey Planellas[3], Giulia Micallef[4], Paulito Palmes[1], Chris Webb[5], Ramon Filgueira[6] and Jon Grant[2]



Visualisation of AutoAI results interrogating fish response



github.com/iBM/precisionAquaculture.jl/

Shellfish farm closure

- Water quality events can lead to regulatory closures of shellfish sites to prevent contamination
- We have anecdotal knowledge of conditions that lead to closures (algal blooms, nutrient upwelling, urban runoff, etc.)
- Can we use machine learning to forecast shellfish site closures and use that model to interrogate drivers.
- Pragmatic approach: The science is complex but early-warning can be sufficient (without the biology)
- Focus on two sites in Portugal and Northern Ireland – incorporate local conditions and site variations.

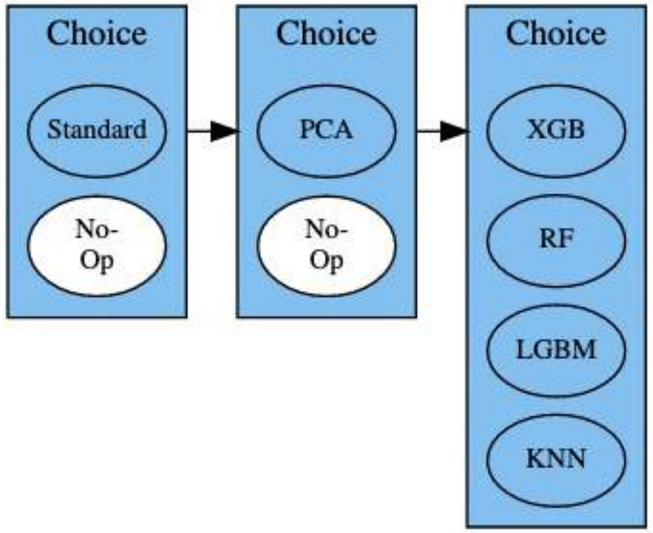
Data driven exploration

The Data

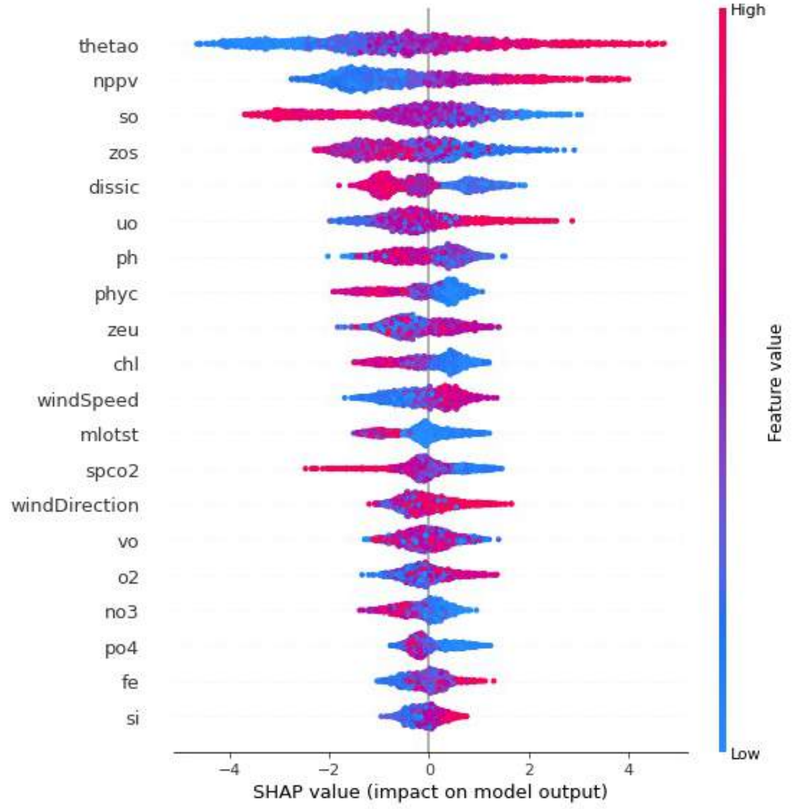
- Historic site closures
- In-situ env. sensors
- Satellite and open-ocean data
- Weather data
- *Waterbody attributes* – closure conditions vary hugely across sites

Machine Learning

- Semi-automated ML to interrogate the solution space



Improve AI explainability with IBM Cloud Pak for Data



Conclusions

- Standardize – data streams, semantics, external data
- Automate – data integration, data quality, model development, model predictions
- Customize – enterprise operations, regulatory requirements, geographical considerations.
- Decide – probabilistic forecasting and risk management, human-in-the-loop resources,