This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 773330

Deliverable report for

GAIN

Green Aquaculture Intensification
Grant Agreement Number 773330

Deliverable D.1.3
Title: Methodologies for big data mining in aquaculture

Due date of deliverable: 30/04/2019
Actual submission date: 30/04/2019

Lead beneficiary: IBM
Authors: Fearghal O'Donncha, Mark Purcell
WP: 1 Production and Environment
Task 1.4: Development of a real time Information Management System

Dissemination Level:		
PU	**Public**	**Y**
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Document log

Version	Date	Comments	Author(s)
Version 1	28/03/2019	Table of contents	Fearghal O'Donncha
Version 2	10/04/2019	First draft	Fearghal O'Donncha
Version 3	19/04/2019	Reviewer's comments	Roberto Pastres, Andre Lopes
Version 4	26/04/2019	Address reviewer comments	Fearghal O'Donncha

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

Version 5	29/04/2019	Reviewer comments 2	Joao G. Ferreira
Version 6	29/04/2019	Final version	Fearghal O'Donncha

Recommended Citation

O'Donncha, F., Purcell, M. Methodologies for big data mining in aquaculture. Deliverable 1.3 GAIN - Green Aquaculture INtensification in Europe. EU Horizon 2020 project grant nº. 77330. 1-28

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

GLOSSARY OF ACRONYMS

Acronym	Definition
ANN	Artificial Neural Networks
Chl a	Chlorophyll a
CSV	Comma Separated Value
DO	Dissolved Oxygen
ECMWF	European Centre for Medium-Range Weather Forecasts
FCR	Feed Conversion Ratio
GAM	Generative Additive Models
IoT	Internet of Things
LSTM	Long-short-term memory
MAPE	Mean Average Percentage Error
ML/DL	Machine Learning/Deep Learning
MLP	Multilayer Perceptron
MQTT	Message Queue Telemetry Transport
PAIRS	Physical Analytics Integrated Repository and Services
PFF	Precision Fish Farming
SST	Sea Surface Temperature
RF	Random Forest
TWC	The Weather Company
XGBoost	Extreme Gradient Boosting

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

Table of Contents

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

1. Introduction

Many scientific and commercial applications require the ability to infer meaning from data and take action based on that meaning. Fundamental to value extraction is curation and analysis, with rapid, near real-time processing often a key requirement (Hey, Tansley, & Tolle, 2009). These steps demand the ability to process, interpret, cross-correlate and integrate data from a wide variety of heterogenous sources in a scalable manner, and return information from the data to the stakeholder in actionable form.

Data generated on modern aquaculture farms extend across a wide variety of forms. In situ sensors sample large numbers of environmental variables such as temperature, current velocity, dissolved oxygen (DO), chlorophyll and salinity, and can be considered timeseries data (as it often returns data at a single point or set of interconnected points). Remotely-sensed environmental data can sample much larger spatial domains and can be at the bay-scale – from land-based sensors such as CODAR-type HF radar (O'Donncha, Hartnett, Nash, Ren, & Ragnoli, 2015) – or at the global scale from satellite monitoring systems (von Schuckmann et al., 2018). Monitoring of farm operations also requires sampling of animal variables (Føre et al., 2018) such as size, clustering behaviour, and movement, and this is typically done using underwater technologies such as video monitoring, hydroacoustic technology and aerial drone imagery. Further, there are large datasets of pertinent variables that are generated by numerical models such as weather or ocean circulation products. These datasets constitute huge data volumes with distinct characteristics. Integrating and extracting information from these data sources are key to encapsulating the full dynamics of the farm environment and enabling effective farm management.

A key objective of the GAIN project is to "develop and test precision fish farming (PFF) for management of finfish and shellfish farms by combining sensors, key performance indicators, big data analysis, and predictive mathematical models for production, animal welfare, and environmental effects". This deliverable describes the role of data to enable PFF and details an effective strategy to curate and exploit these data across all levels.

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

2. Motivation and requirements analysis

Data generation, curation, analysis and enhancement involved multiple partners:

- Ten pilot study sites were instrumented with a large number of sensors (see GAIN Deliverable 1.1 (Service et al., 2019) for details) and, using a range of technologies (mobile telemetry, Wi-Fi, etc.), data was transferred to the cloud for further processing.
- Actionable insights from data rest on the availability of the pertinent feature variables (measurable property that is relevant to the target variable we wish to optimise, e.g. temperature, dissolved oxygen) and target variables (property we wish to optimise such as farm productivity or feeding rates) to guide decision making. The GAIN consortium provided domain expertise and led the development of mechanistic models that augment data being collected (e.g. models of fish growth).
- IBM developed a cloud deployment that will integrate the disparate datasets being generated across all volume scales onto a single platform that enables integrated analysis of datasets from all sources described above (timeseries, satellite, acoustic, video, numerical model). These are complemented by a suite of data-driven models to move data towards actionable insight and decision.

In the first 6 months of the GAIN project, all partners participated in a detailed analysis of data currently being generated across various sources, additional data required to accurately monitor pertinent events and processes, and the key "burning questions" facing the farm operators that PFF could help alleviate.

In September 2018, a GAIN partner meeting was hosted on the PFF topic, attended by a number of the industry, scientific and technological partners. From that meeting, partners outlined the preliminary requirements and objectives from PFF. The main focus of the meeting was:

- "Burning questions" facing the aquaculture industry and how technology and science can address these questions and add value to the industry.
- Sensor data collected at farm sites – data type, volume, veracity, resilience, communication and scientific and operational value were assessed.
- Data-driven and mechanistic models that will be implemented during the project – predictive variables, model capabilities and requirements (training data both historical real-time, operational data on farm management, production data).
- Integration of the different components – sensors data to real-time platform, models and data, unstructured data with predictive models, predictive models with operational decision.

The main outcomes from the meeting were:

- IBM needs inputs (explanatory variables) and outputs (response variable) for their semantic model. Fundamentally, the objective is to relate explanatory variables (temperature, DO, water currents, etc.) with response variable (fish growth/health,

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

farm productivity, etc), that will allow for these response variables to be predicted. We agreed that inputs would be environmental parameters collected by the real time sensors (temperature and DO for salmon farming, possibility of including salinity for Scotland), satellite data (including historical data). Data from operational, open ocean model forecasting tools (also historical data) and potentially mechanistic models developed by the consortium can potentially be included here. Outputs will be the behavioural data (translated to quantitative information on fish status) recorded in real time by the sonar system deployed (CageEye or other), together with other quantitative information on farm production (fish growth, fish health, yield, etc.). Behavioural data will have to be interpreted, quantified (frequency of occurrence per day for example) and scored as positive or negative when correlated with the stressor. A further validation with physical sampling, management and production data (sea lice counts, gill heath, mortalities and FCR (Feed Conversion Ratio), or yield for example), provided by the farmer would be ideal but this will have to be negotiated with industry partners. Instrumentation of pilot farm sites was designed to collect environmental and behavioural variables to best describe bay processes and farm conditions. Full details on these datasets are provided in Deliverable 1.1 (Service et al., 2019)

- The University of Dalhousie provided a list of burning questions facing the industry to improve the health and welfare of salmon, and for production purposes. Examples of these included

    1. What is the scale of variation in oxygen and temperature – cage, farm, regional, vertical, temporal?
    2. How is fish health/welfare affected by weather, oxygen, feeding, disease etc.?
    3. How can trends in oxygen and temperature be predicted over several days to a week?
    4. How effective is supplemental oxygenation, and how can it be optimized

    The project will have to try to address these questions by using PFF.

- IBM will run a pilot test of their model using actual data on sensors already deployed and also based on historical data. The objective is to show the stakeholders involved in the project the use and applications of the model on real fish production systems

The motivations, detailing and descriptions of the datasets being generated were comprehensively described in Deliverable 1.1 (Service et al., 2019).

3. Methodology

Big data analysis in aquaculture has several primary considerations:

- Large volumes of Internet of Things (IoT) datasets curated in a manner that is scalable and rapidly queryable.
- Machine learning-based analysis that can augment and extract insights from data and can be readily adjusted to other study sites.

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

- Cloud environment to provide scalable computational resources to stakeholders.
- Insights and decision extracted from data using data-driven automated approaches.

This section presents the approach adopted in GAIN to address these considerations.

3.1 Data generation and curation

As mentioned previously, aquaculture data come from a large variety of sources, including environmental sensors, underwater video monitoring, hydroacoustic technology and drone imagery, together with satellite and other geospatial datasets. Effective management and curation of those datasets require consideration of the temporal, spatial and volume scales.

Large volumes of data generated in aquaculture consist of time-series data which return multiple measurements in time at specific coordinates in space. These data often 1) are collected in harsh environments from in situ sensors and 2) have a very limited shelf life, before the value of the data begins to decay. A key consideration with these datatypes is transfer through low bandwidth connectivity to services platform for further processing and integration.

A related dataset is that which extends over large spatial domains and can be termed geospatial data – that is, data with location information – to differentiate the point source data (which often also has location information but for convenience, we differentiate), described above. Common sources of geospatial data pertinent to aquaculture include satellite data, drone imagery, and weather and other numerical model generated data. The primary characteristic of these datasets is huge volumes – often in the hundreds of terabytes. A key challenge with geospatial datasets is rapid querying due to the large volumes, often in heterogeneous formats with inefficient data access patterns.

Integration of these two datatypes with a services platform has very different characteristics, namely: 1) a simple, low-bandwidth approach to communicate from farm to the data platform and 2) an approach that enables scalable curation and rapid querying of the huge datasets available from geospatial sources. We will describe the system that enables the management and enhancement of these datasets

3.1.1 Geospatial data

Geospatial datasets pertinent to aquaculture, include sea surface temperature (SST), chlorophyll a (Chl a) and ocean waves, that are monitored daily from multiple satellites circling the globe. In addition, weather, ocean, and climate models compute the current system state multiple times per day, enabling insight into atmosphere-ocean interactions as well as providing means to monitor and predict macro-scale events on the bay- or farm-scale. Traditionally, these datasets have not been fully exploited by aquaculture stakeholders due to the huge volumes and the difficulty in querying data from heterogeneous file, as well as concerns over nearshore accuracy and spatial resolution. In the geosciences, netCDF, HDF and GeoTIFF are some of the most widely used data storage formats. These are fundamentally based on storing and retrieving data in the form of arrays with associated meta-descriptors. While this is an efficient storage method, it makes querying of specific datasets – and in

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

particular querying relationships between multiple datasets – cumbersome if not impossible, without loading and acting on entire files.

The solution to allowing rapid analysis and processing of satellite and geoscience data (before the value begins to decay) is to have all data layers curated and homogenized before being uploaded to the platform, eliminating the time required for data pre-processing. Data curation require data validation, verification, and alignments spatially and temporarily, such that these layers are ready to be integrated into physical and machine learning models without the need for data download, validation, and pre-processing. Traditional database technologies do not allow efficient indexing and joining of data layers once data volumes exceed a few terabytes.

IBM research has previously investigated the problem of processing geospatial datasets applied to the agriculture industry (Mello & Raghavan, 2018). The work resulted in a geospatial big data platform, Physical Analytics Integrated Repository and Services (PAIRS), that aims to process petabytes of data and address the spatial and temporal complexity associated with heterogeneous data integration (Klein et al., 2015).

The PAIRS architecture automatically downloads, curates and stores real-time geospatial datasets in a scalable storage table, which are then available for ingestion in real time data-driven and mechanistic modelling systems. PAIRS offers access to a repository of consistent historical and real-time datasets that are aligned and indexed. Data curation encompasses conversion to a common datum aligned on a well-defined spatial grid. The platform can be queried to retrieve data in multiple ways: (1) single point across large interval to create time series, (2) spatial query across an arbitrary sized area, and (3) filtered spatial and temporal query using a system of filters to retrieve subset of data from each layer.

It is developed on top of the open source big data technologies Hadoop and HBase (George, 2011). It leverages MapReduce (Dean & Ghemawat, 2008) to accelerate data queries by parallelizing search and data retrieval. One key differentiator of PAIRS is the multi-layer query capability, ability to search multiple data layers and filter then based on multiple search criteria where filters allows discovering locations or time periods that share the same characteristics in space and time. This capability provides a quick way to visualize in real time changes for a certain location and detect similarities or differences. Technical details on the PAIRS architecture are provided in (Klein et al., 2015).

As part of the GAIN project, we will extend PAIRS with aquaculture-specific capabilities. In particular, we will add ocean-related features and the ability to query, filter and download the data at user-defined locations across large time intervals. This will drastically increase accessibility to geospatial datasets for aquaculture industry stakeholders (operators, academia, regulatory), serving to unlock this big data potential.

3.1.2 Timeseries data

Timeseries data generated on the farms are, usually, in terms of date-stamped comma-separated value (CSV) files. For the different sensor configurations considered, we develop a code that is deployed by the users (pilot study partners). It converts the sensor data from the raw CSV format into a messaging, MQTT format that is compatible with standard IoT platforms

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

and periodically pushes the sensor data to the IBM platform.

The public MQTT sample code available at https://github.com/GoFlexH2020/samples provides a template for partners to submit their sensor data to the Service Platform. MQTT (Message Queue Telemetry Transport) is an ISO standard for lightweight publish-subscribe style messaging built on the TCP/IP protocol. It is designed for remote locations, and low power devices which are network bandwidth limited. The sample code highlights the principal steps to pre-process data prior to submission: conversion of timestamps to UTC, anonymization of data (if appropriate), and structuring of a valid message. Request timestamps are in ISO8601 extended format as shown and must be UTC aligned (`YYYY-MM-DDThh:mm:ss+00:00`).

The public MQTT sample code is written in Python. Python was used as it is simple to read, and generally understandable even by non-programmers. It provides the following files:

- `README.md`
- `goflexsubmitapi.py`
- `messaging.pem`
- `mqtt_client.py`
- `requirements.txt`
- `sample.csv`

`goflexsubmitapi.py` contains a Python class `GoFlexMeterSubmissionAPI` which encapsulates and simplifies writing an MQTT client, reducing the process to a small subset of functions.

As the security and integrity of the data in transit are important, the provided class also demonstrates how to leverage the authentication and security options of the Service Platform data ingestion module (IoT Platform) using a combination of SSL certificates and tokens. mqtt_client.py demonstrates how to use the `GoFlexMeterSubmissionAPI` class as well as providing function stubs for the expected data pre-processing steps such as data anonymization. Data anonymization can become an extremely complex task and is beyond the scope of this sample code. Appropriate anonymization techniques are specific to each data provider, where there may be different regional regulations. It is the responsibility of each demonstration site to apply adequate anonymization to their data prior to submission to the Service Platform.

3.2 IoT platform

The IoT Platform on IBM® Bluemix™ provides a single point of ingress for all sensor data generated for the GAIN project. It utilizes the MQTT Protocol for data transport. The service also provides secure transport to ensure the integrity of the data in transit and token-based authentication to restrict data submission to known devices. It is designed to scale to thousands of devices in parallel, with each device providing periodic or intermittent updates. The service can be geolocated in any of the Bluemix™ supported regions (US, Germany, UK,

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

Australia). For GAIN, since it deals with European data, the service is currently located in the UK. In fact, an instance of the service is created for each partner site. The IoT platform makes no provision for data storage, but can readily connect to additional Bluemix™ storage services.

Sensor data sent to the service is in the form of short JSON strings – a lightweight data interchange format. Creating the service is achieved on either the Bluemix™ Dashboard (web interface) or using the command line client. In the dashboard, a search for "Internet of things platform" displays the service. See Figure 1, click the icon, populate the page, with details such as "Service Name", preferred region and click create, see Figure 2. Once the service is created, device types, devices, and authentication tokens for the devices can be generated.

An additional feature of the IoT service is the ability to interrogate data from each device as it arrives in real time. As a device submits a reading it is available on the dashboard, see Figure 3. This is useful for monitoring and is especially useful for debugging potential client submission messages as well as end to end workflows.

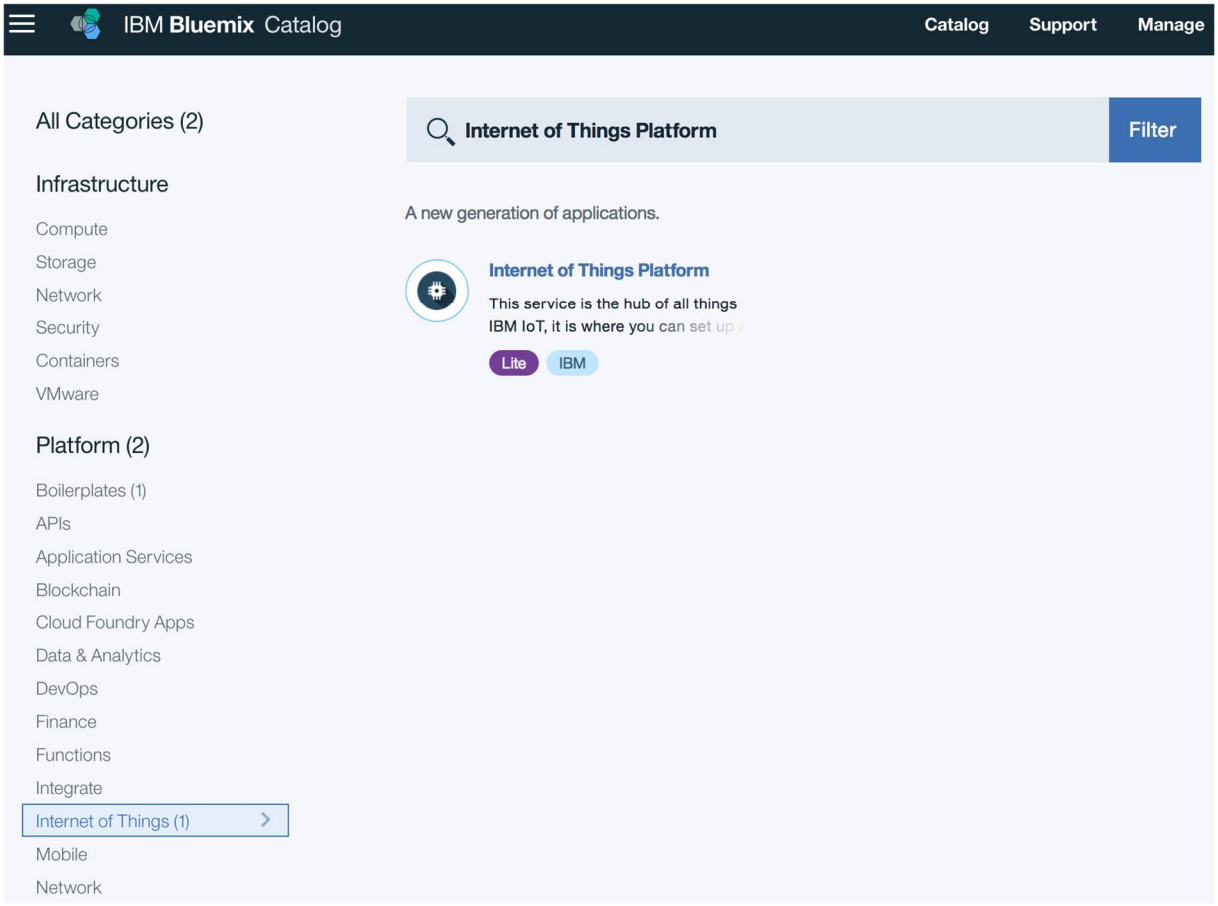

Figure 1: Choosing the IoT Platform from the Bluemix™ web console

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

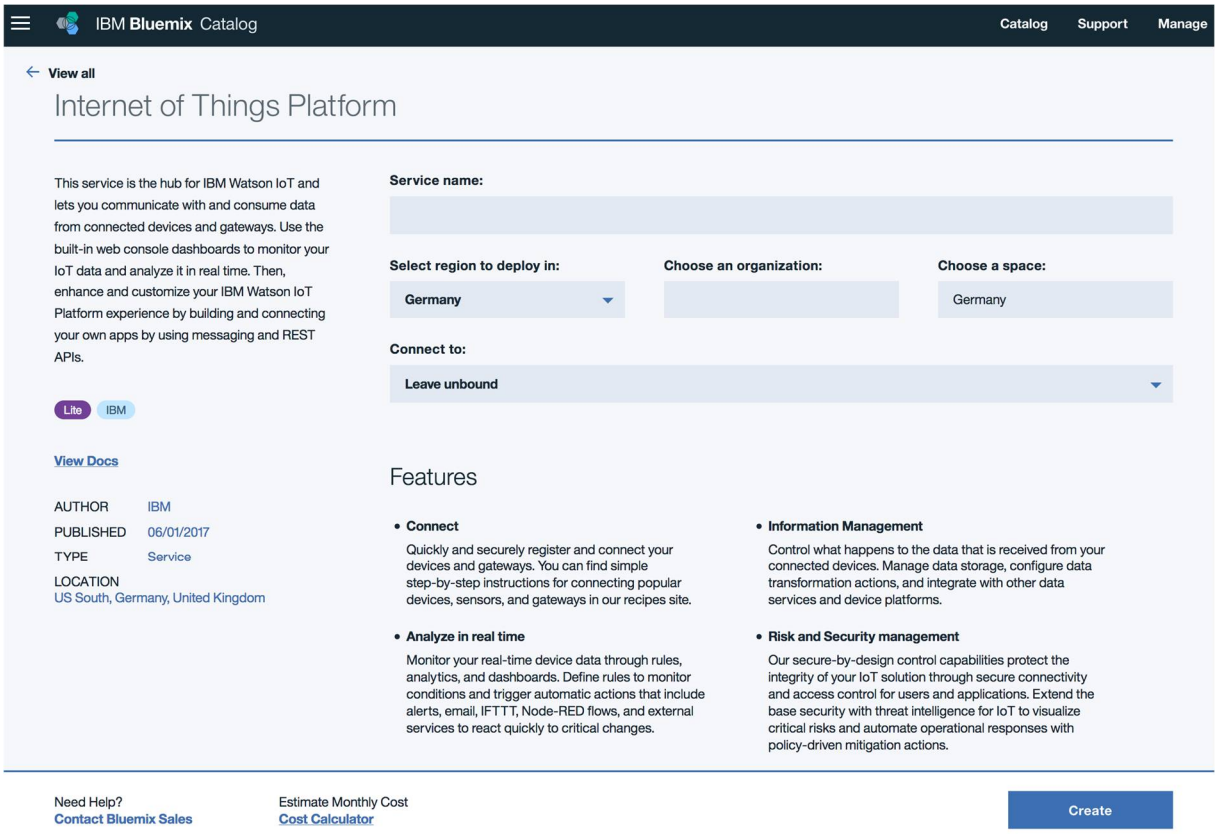

Figure 2: Creating an instance of the IoT Platform on Bluemix™

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

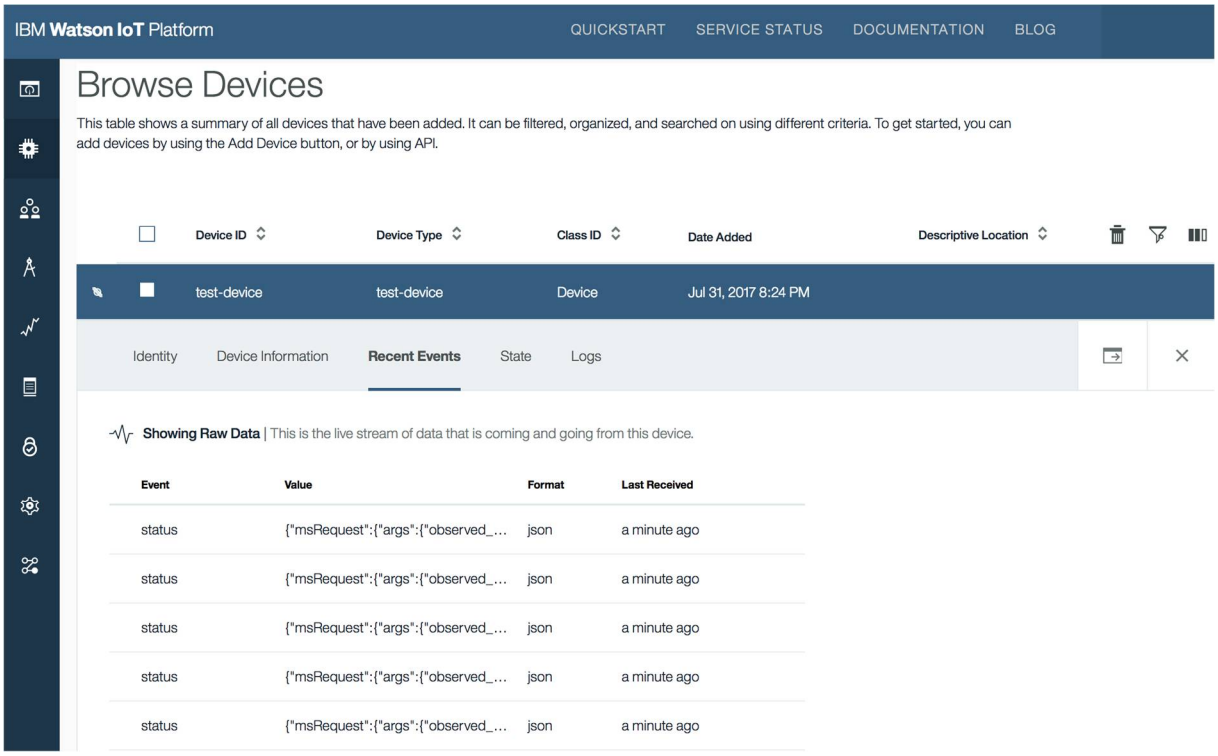

Figure 3: IoT Platform dashboard, showing some incoming test data

3.2.1 High Throughput Messaging System

IBM Message Hub is a scalable, high throughput message bus. It can be used to wire micro-services together in order to stream continuous data from one service to another. It is designed to handle real time data feeds such as environmental sensor data collected in GAIN. Its queues (identified by a topic name) also have a user-defined persistence duration, which provides a buffer to hold data in the event of a micro service failure. As stated previously the IoT Platform does not provide storage. Message Hub is an ideal candidate to receive data from the IoT Platform as it allows data to continue flowing through the Service Platform whilst providing a temporary buffer should anything fail.

3.2.2 Timeseries storage

As the sensor data for GAIN is structured data, an SQL style data store is an obvious candidate for long term storage. It allows for the creation of table indexes for fast lookup of ranges or by sensor type. It also allows for the creation of stored procedures, which simplifies client insertion code by freeing it from having to know about the underlying table structure. IBM® DB2® on Cloud is a fully managed SQL database, with several client runtimes supported. It is available on Bluemix™ and integrates readily to existing micro-services. It also provides a dashboard for handling and querying the stored data. Data is also accessible via a number of additional methods from command line client to REST interface.

3.3 Service description

The Service Platform is a fully cloud hosted, cloud native system, based on a micro-services

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

architecture (Newman, 2015). The cloud infrastructure is provided by IBM, using the IBM® Bluemix™ platform.

The micro-service interactions interface is based on the Publish/Subscribe Design Pattern (Tarkoma, 2012). Each micro-service either subscribes to events of interest, publishes information for subsequent consumption, or both. In either case, the content of information published or received (via subscription) is called a message which are in JSON format.

Employing this design pattern enables asynchronous interactions between components, whereby a component can publish several messages in quick succession, and then subscribe to generated events (if appropriate).

An example of this is during the training of machine learning models to predict a relevant variable, e.g. SST. This requires large volumes of historical SST data (e.g. from in situ sensor or from satellite data via PAIRS) and matching weather data. As separate micro-services handle the different requests, these can effectively operate in parallel. The model training micro-service then subscribes to the relevant output and waits for both sets of data. Different micro-services may take varying amounts of time to complete their requested workload, and as such, the data returned may arrive out-of-order at the model training micro-service. This out-of-order arrival is a natural consequence of asynchronous systems and must be handled appropriately by the requesting micro-service which is supported at the message-level.

3.4 Model development

A key functionality of the service platform used in GAIN is the ability to manage different models and integrate with the different data streams coming from farms, PAIRS, weather data, and other open sources. Models are trained using historical features and labels in machine learning parlance. Features include variables such as weather data, nutrients, currents, ocean temperature, etc. while labels include variables such as fish behaviour, growth as well as environmental variables that one wishes to predict, such as water temperature or Chl a.

Training service follows the sequence of events mentioned below:

1. The training function is triggered by a model training request message.
2. Historical feature data and label data as payload is published in the triggered call.
3. Training micro-service subscribes to an AMQP queue and gets the payload from the queue.
4. Training micro-service trains the model using the payload by running a Docker container and returns a data frame of the trained model.

Within the context of the GAIN project, a wide number of different models are developed, focusing on different components of operations. Depending on the specific application, various models across the statistical, machine learning (ML) and deep learning (DL) spectrum are adopted.

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

Models are developed using the Python language, although other languages such as the R (Team, 2013) framework are also supported. The implementation of these models within the IBM service platform involves the following phases

1. Load data: Loads the data provided as payload from AMQP model into Python and converts it into a dataframe.
2. Data curation/manipulation: Covariates and any additional column vectors that can influence variable to be predicted are defined from the payload data dataframe. Any discrepancies in the data are cleared in this step as well.
3. Data training: Model is trained on the available features and labels until the model can accurately predict the selected variable. The trained model is returned to the AMQP queue.
4. Model Scoring: The trained model is packaged as a Docker image managed by the IBM service platform. It can be called to make prediction against future state as new data comes in.

4. Data-driven model development in GAIN

In this section, we describe development and performance of a number of data-driven models pertinent to aquaculture operations. These demonstrate the avenue opened up by the big data described in the previous section: using big data to train machine-learning-based forecasting models. Once trained, the computational expense of these products is low and, conceptually, one can develop transferrable models that can be trained to learn features at different geographical location. Here we consider SST as a representative example relevant for aquaculture operations from a number of perspectives.

4.1 Statistical and machine learning approaches to forecast sea surface temperature

Sea surface temperature (SST) is a common indicator of biological productivity in aquaculture (Handeland, Imsland, & Stefansson, 2008) and has known effects on fish behaviour. Currently, most forecasting systems for SST are based on the solution of the heat transport equation coupled with ocean (or coastal ocean) models that resolve the Navier-Stokes equations. These are often integrated with data assimilation schemes that update the accuracy of forecasts with observations using a model-data filtering approach. The current state-of-the-art is the ERA5 global dataset from the European Centre for Medium-Range Weather Forecasts (ECMWF) (Hirahara, Balmaseda, & Boisseson, 2016). It provides short-term SST forecasts (and hindcasts) on a 32km horizontal grid at hourly intervals from a numerical synthesis of ocean models, atmospheric forcing fluxes, and SST measurements.

These analyses and forecasting systems face a number of scientific, technical, and practical challenges. The computational and operational requirements for ocean simulations at appropriate scales are immense and require high performance computing (HPC) facilities to provide forecasts and services in practical time frames (Bell et al., 2015). Operational forecasting systems require sophisticated data assimilation schemes that takes account of bias

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

and errors in models and observations. Proposed schemes consider a wide variety of approaches to account for different error sources broadly decomposed across *a-priori* based approaches that incorporate information on physical relationship between variables and *statistical* approaches that are typically based on ensemble model projections (Martin et al., 2015). The ERA5 reanalysis uses a three-dimensional variational DA scheme (3D-Var). A consequence of these challenges is that operational forecasting systems are only feasible for large research centres or collaborations who have access to large-scale compute resources and scientific expertise.

Within the GAIN project, we will develop a number of data-driven models to fully exploit the potential of the data being collected and improve farm operation using a set of highly localised monitoring and predictive components to fully inform on farm dynamics. We will briefly describe the different models developed to forecast SST.

4.1.1 SST forecasting models

We developed a set of predictive models, namely, GAM, RF, XGBoost, MLP, and LSTM to forecast SST. Combining different models increases the robustness of the approach as no single model is expected to perform best across all conditions based on the well-known, 'no free lunch' theorem (Wolpert & Macready, 1997). The objective of each selected model was to relate a univariate response variable **y** to a set of explanatory variables **x** = x_1, x_2, ... ,x_n (representing for example, historical SST, air temperature, seasonal identifier, etc.).

Generative Additive Models (GAM) characterize general nonlinear regressions without requiring pre-specification of the form of the nonlinear relationship (Hastie & Tibshirani, 1986). Broadly, the method forecasts the response variable (here SST) by taking into account the covariates that can influence it (e.g. wind speed, air temperature, solar radiation, cloud cover) and represent them as splines of varying degrees of freedom. The degrees of freedom of these splines are computed using the historical data.

A Random Forest (RF) is an ensemble of decision trees 'grown' from a subset of the training data. Each new training set is drawn, with replacement, from the original training set. Then a decision tree is grown from the new training set using random feature selection (Breiman, 2001). Outputs of all trees are averaged to provide the prediction. Some of the key advantages of RFs are efficiency (easily parallelised), robustness to outliers and noise, and insight to feature importance during model interpretation.

Extreme Gradient Boosting (XGBoost) shares many characteristics with RFs, with a key difference being that decision trees are built *sequentially* rather than *independently*. In this manner, subsequent decision trees are built to further minimise prediction residuals (thereby theoretically improving accuracy). Amenable towards parallelisation, XGBoost provides a computationally efficient, high performance decision tree implementation that has shown impressive success rates in applied machine learning application and competitions (Chen & Guestrin, 2016).

From the deep learning (DL) library, we first implemented a multi-layer perceptron (MLP) model. The MLP conceptual model is a supervised learning approach that is loosely based on

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

the anatomy of the brain. Such an artificial neural network (ANN) is composed of densely interconnected information-processing nodes organized into layers. The connections between nodes are assigned "weights," which determine how much a given node's output will contribute to the next node's computation. During training, where the network is presented with training data (i.e., the response variable (SST) at each point in time and associated covariates (air temperature, solar radiation, wind speeds, etc.)), those weights are optimized until the output of the network's last layer consistently approximates the result of the training data set (in this case, SST).

As shown in Figure 4, an MLP model is organized in sequential layers made up of interconnected neurons. By minimizing the loss function, the supervised machine learning algorithm identifies the hyperparameters (values of weights and biases assigned to each "node") that yields a predicted value, $\hat{y}$ that matches the measured value, y (Goodfellow, Bengio, & Courville, 2016). As shown in Figure 4, a machine learning model transforms an input vector (layer) to an output layer through a number of hidden layers. The machine learning model is trained on a data set to establish the weights parameterizing the space of nonlinear functions mapping from x (covariates) to y (response variable). Of course, a large training data set is required to develop a robust machine learning model: one of the key objectives of the GAIN project is to fully exploit the value of big data being generated on modern aquaculture farms to enable, develop and exploit deep learning approaches.

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

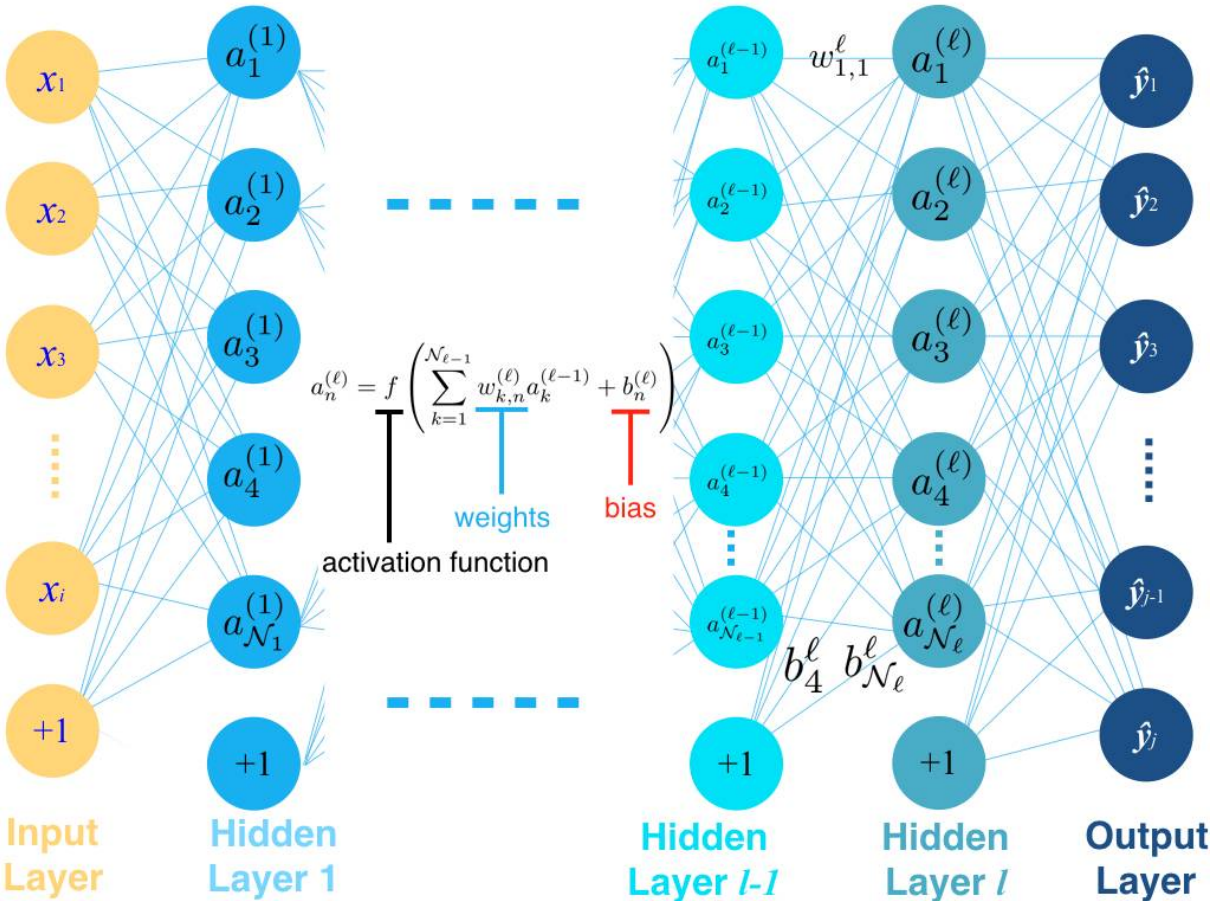

Figure 4: Schematic of an MLP machine learning model

It is intuitive that SST has a recurrent component (i.e. the temperature today has some dependence on temperature yesterday). Traditional artificial neural networks (ANN) consider each time point independently and do not have any "memory" component (i.e. conditions at previous time). An alternative formulation that has demonstrated huge success in time series forecasting implementations such as precipitation (Xingjian et al., 2015), traffic forecasting (Zhao, Chen, Wu, Chen, & Liu, 2017) and natural language processing (Wen et al., 2015) is long-short term memory (LSTM) models.

A fundamental extension of LSTM compared to ANNs is parameter sharing across different parts of the model. A LSTM with a single cell recursively computes the hidden vector sequence **h** and output vector sequence **y**, recursively thereby enabling model parameters to be updated with information at preceding times. This parameter sharing allows the model to incorporate information from previous timesteps when making a prediction for the current timestep.

4.1.2 Application to satellite data

Training data for this study were the MODIS instrument aboard the NASA *Aqua* satellite. MODIS SSTs are produced and made available to the public by the NASA GFSC Ocean Biology Processing Group. The MODIS sensor measures ocean temperature (along with other ocean products such as salinity and chlorophyll concentration) from a layer less than 1mm thick at

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

the sea surface. Data are available from 2002 to present at 4km horizontal resolution and daily intervals. Calibration of the Pathfinder algorithm coefficients and tuning of instrument configurations produce accurate measurements of SST with mean squared error (MSE) against *in situ* sensors of 0.2°C (Kilpatrick et al., 2015). These accurate global SST measurement on decadal periods, serve as an ideal dataset to extract insights using ML.

Weather data also serve as features to the models and were extracted at 30-km spacing from The Weather Company (TWC). Hourly forecast data out to ~15 days are available along with historical cleansed (i.e., subject to quality assurance procedures) data for the past 40 years on a regular grid.

Initially, we deployed the models at an arbitrary mid-ocean location using a train/test approach; the model architecture (number of splines for GAM, number of trees for RF, number of nodes and layers for DL models) was parametrised at this location and performance evaluated against measurements (test data). Satellite measurements from MODIS spectrometer were collected over 16 years from July 2002 to December 2018. Optimising model performance included both feature selection and parameter tuning. In machine learning, feature selection relates to the specification of appropriate covariates to explain the response variable. Some of the features included in this study were weather data, day of year and autoregressive (i.e. SST at earlier time steps) data. Parameter tuning is part of model training in machine learning, where the model is fed large volumes of data until the optimal set of parameters are computed, typically by some type of gradient descent. In this study, parameter optimisations adopted a greedy, grid-search approach over user-defined parameter ranges. 90% of the data was used to train the models and 10% was held back as test data to evaluate performance. Model performance were evaluated based on the mean average percentage error (MAPE) compared to satellite observations. MAPE has been shown to be an effective measure of quality for regression models (de Myttenaere, Golden, Le Grand, & Rossi, 2016).

Figure 5 compares model predictions for the test period to MODIS data. Observing the time evolution of SST reveals that a suitable model must represent two distinct time scale components. On the one hand, there is the smooth SST evolution governed by seasonality. This component of SST evolution benefited from suppression of large fluctuations. Of the models studied, this criterion was fulfilled by the GAM approach, which yielded lowest test MAPE with a large regularization parameter (i.e., the penalty on the second-order derivative of fitted single-feature functions). On the other hand, the seasonal cycle has superimposed on it, short-term behaviour dominated by peak events occurring at daily to weekly time scales. This is particularly evident in the XGBoost and the MLP approaches where both yielded best performances for smaller regularization terms, which enabled them to better capture short-term events.

The large regularization effect together with the piecewise polynomial components of GAM models contributed to a “smoother” time-series prediction that still captured long-term trends, including correlation of data between years. Similarly, the RF approach led to a comparably smooth SST evolution but at significantly lower MAPE than the GAM model. The most obvious reflection of the seasonal trend is evident in the LSTM prediction which

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

produces a highly smoothed representation of the training data. The model fails to capture any small-scale dynamics at the daily or weekly level instead reproducing the seasonal heating/cooling effects only. Further analysis of model parameters suggested this to be a result of the retained long-term memory informing the broader trend only It's worth noting that while XGBoost and MLP captured the small-scale fluctuations better, RF returned lowest MAPE. To simultaneously take both aspects into account, a model combination weighted by the performance of individual models can serve to better capture varying dynamics and improve predictive skill.

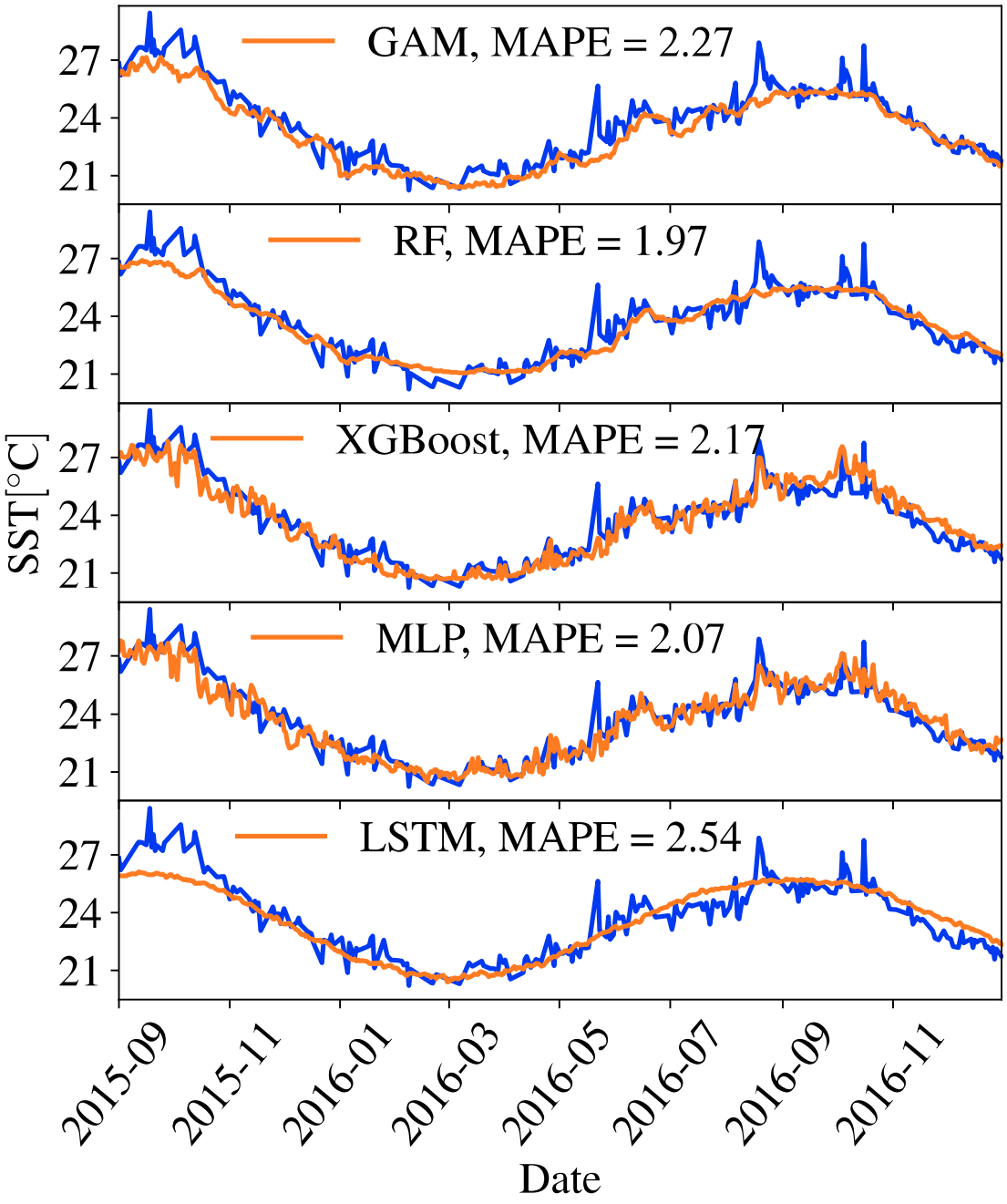

Figure 5: Sixteen-month forecast (orange curve) of SST at location A from different ML models trained on 12

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

years of preceding historical data compared to MODIS measured SSTs (blue curve).

4.1.3 Ensemble aggregation and transferability

As the feature engineering (selection of covariates that influence response variable, either manually – based on domain expertise – or, automatically from statistical analysis), and hyperparameter selection process is complex, it is desirable to execute this procedure at only one location and then use the selected model for retraining at different geographical locations. Further, to exploit information coming from the different models, a method to optimally integrate predictions from different models is desirable.

The underlying assumption of ensemble modelling is that each member contains some information pertinent to the true state of the system. The interplay between models is expected to vary in both space and time; i.e., member models performed better at different points in space and time depending upon ambient conditions, individual model features, and other physical interactions. The objective of the aggregation method was to develop a weight for each member of the ensemble taking into account historic performance of predictions against observations.

Ensemble aggregation techniques can extend from simple arithmetic averages of all models to machine-learning approaches that admit aggregate ensemble predictions based on weighted summation (O'Donncha, Zhang, Chen, & James, 2018). For this study, we implemented a machine learning aggregator that makes use of historical observations and model forecasts to produce a weight for each model (Hoerl, 1985). A linear, convex (i.e., where weights are constrained so they sum to unity) combination of model forecasts is performed with these weights to generate the best model prediction.

To investigate transferability, we applied this approach at two of the GAIN project partner sites – the El Gorguel site operated by Lebeche and located in the South coast of Spain in Mediterranean sea (coordinates 37° 34' 03.211'' N 00° 52' 17.459 W) and the Sagres pilot site located on the Algarve coast, SW Portugal (coordinates 37° 1' N, 8° 53' W). Full details on both sites is provided in Deliverable 1.1 (Service et al., 2019).

The model architecture (i.e. number of splines for GAM, number of hidden layers and nodes for neural network-based approaches, etc.) and features (weather data, historical (autoregressive) measurements of SST and seasonal information), for this study site were equivalent to those computed from the experimental analysis in section 4.1.2. To apply at this new study site, the model simply needs to be retrained on data from this site to update model parameters; the model can then be used for prediction with new incoming data. Since model training is very quick (< 20 seconds for each model), this is of negligible overhead. Training proceeded, as in Section 4.1.2, using a 90%/10% train and test split.

The resulting prediction model was an aggregation of GAM, RF, XGBoost, MLP and LSTM results, where weights were computed for each model based on a minimisation of difference between prediction and measurement (i.e. models that performed better were provided higher weights). Figure 6 compares the ensemble predictions to the satellite measured SST at the two pilot sites.

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

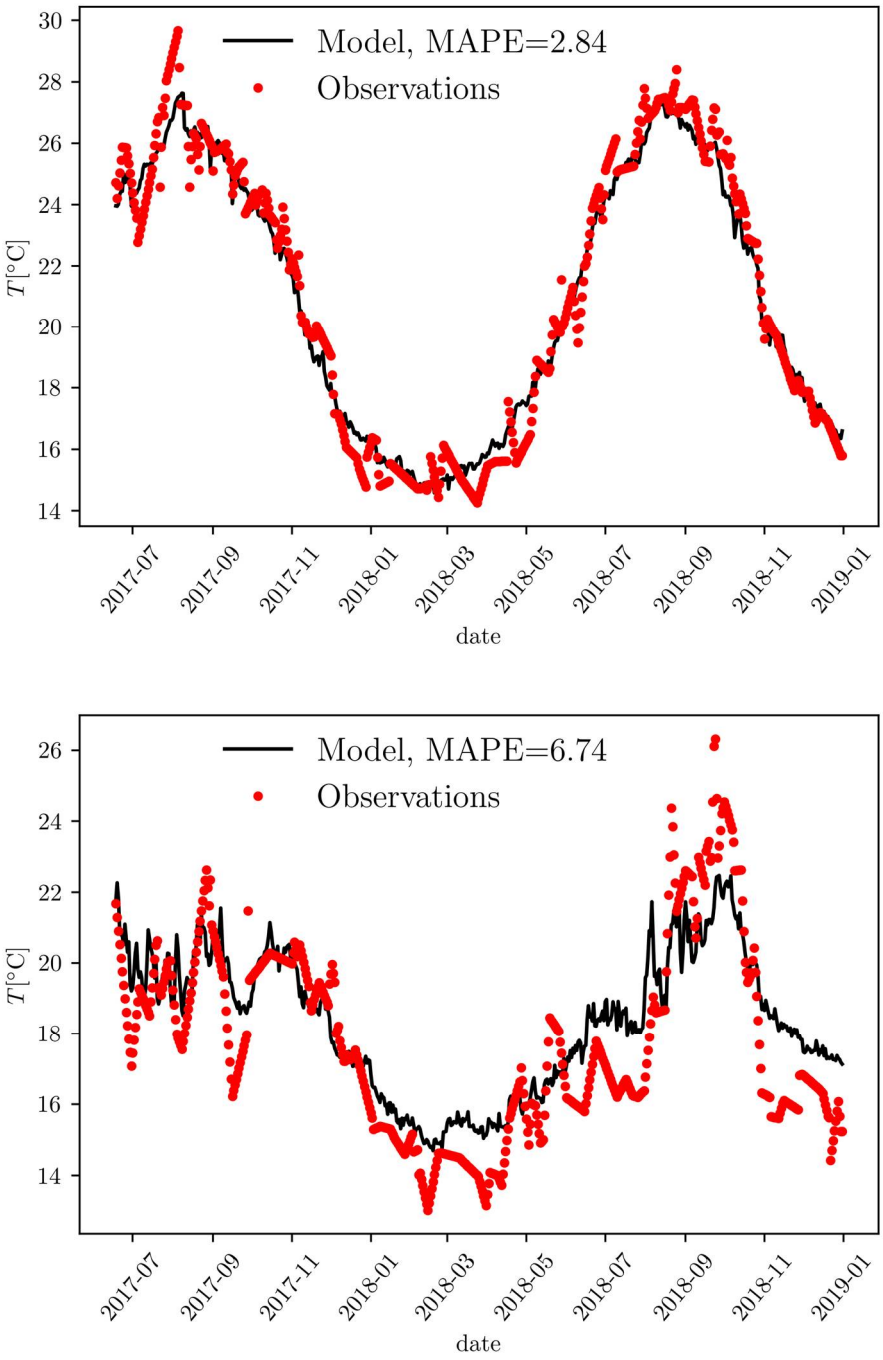

Figure 6: Weighted ensemble average of GAM, RF, XGBoost, and MLP models compared to satellite derived SST at the El Gorguel (top) and Sagres (bottom) pilot farm site

Results demonstrate close agreement with observations, particularly for the El Gorguel site. MAPE over the period is less than 3% and the seasonal trends are closely replicated at this location. The Sagres site exhibits significantly larger temporal variability both for observations and the resulting models. At this site, model approximate measured data well over the first two months, from July to September 2019. Over subsequent time periods, the model tends to drift from observations. This is partly explained by the attempt to perform dynamic forecasts over a 14-month period at a highly dynamic location. In this experimental setup, the model

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

was initialised with observations at time t=0 and for subsequent predictions the autoregressive feature (SST at previous timestep) was replaced by the model computed value (prediction at time t=1 used observation from t=0, while at time t=2 the computed value from t=1 was used). This can plausibly lead to model drift over a 14-month long prediction window.

Figure 7 compares satellite observed SST at the two pilot sites considered. Apparent is the greater variance of seasonal (and shorter time-scale) SST measurements at the Sagres site. At the Lebeche El Gorguel site, SST typically ranges between a minimum and maximum of 13°C and 28°C annually, with these values relatively constant year after year. The Sagres site on the other hand, exhibits significantly higher variance in the minimum and maximum annual values. Maximum annual temperatures vary from approximately 21°C (e.g. 2008) and 26°C (2017) depending on year. This makes long term, predictions based on weather data and seasonal variation much more difficult to capture with data-driven approaches. This is accentuated by the fact that the test period (i.e. period when model is evaluated against measurements), is characterised by higher than average summer temperatures (i.e. in 2017 and 2018).

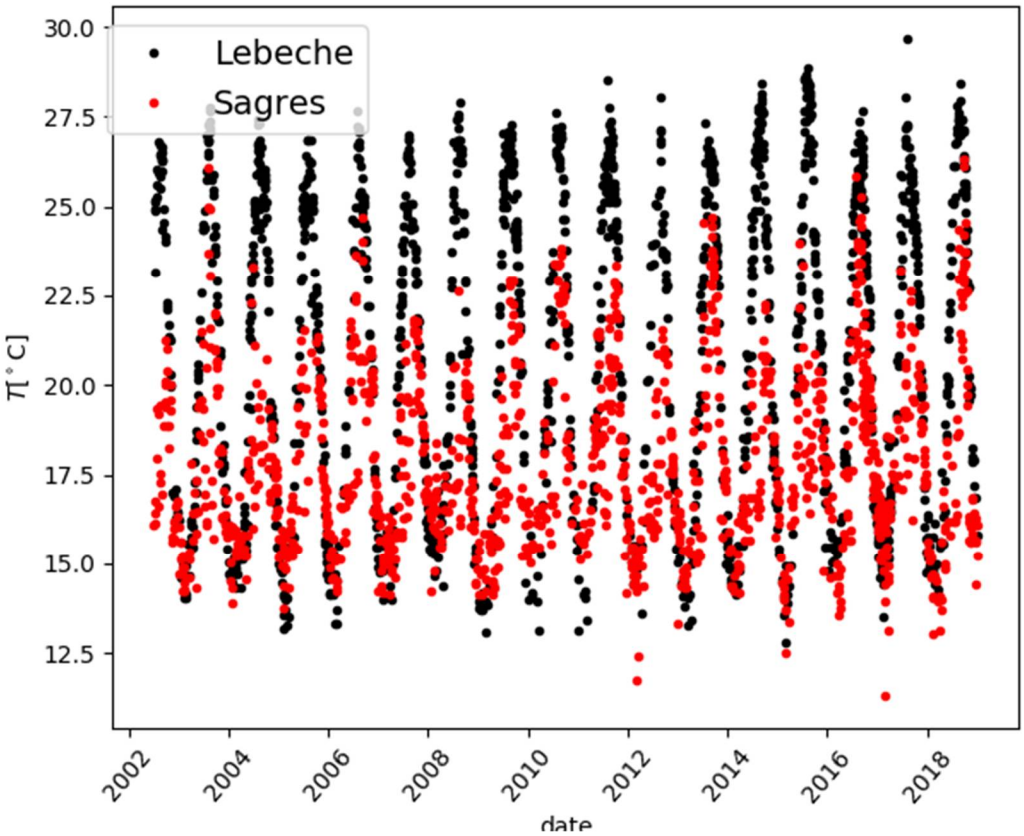

Figure 7: Comparison of satellite observed SST at the two pilot sites over the 16 year study period. The black markers denote measurements at the Lebeche, El Gorguel site, while red denotes measurements at the Sagres pilot site on the Portuguese coast

The higher variance in data at the Sagres site is likely driven by complex phenomena not included as features to the machine learning models such as seasonal upwelling events or land runoff. In a more practical implementation, prediction is limited by the availability of weather

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

data (feature to the model); typically, reliable forecasts of weather can be provided two weeks in advance – this is likely to be a more realistic practical implementation of the SST machine learning model. However, this implementation gives a valuable demonstration of the capabilities of data-driven approaches to predict SST from satellite data at pilot sites.

4.2 Results and discussion

This section considered a framework to develop a transferrable model suite applied to a nonlinear, real-world dataset. Key points considered were design of an automatic feature-engineering module, which, together with a standard hyperparameter optimisation routine, facilitated ready deployment at disparate geographical locations. Results demonstrated that the different models adopted had inherent characteristics that governed accuracy and level of regularization or overfit to training data.

We compared performance of ML models with observations at two GAIN partner pilot sites. We get high Model skill at the Mediterranean location while at the Portuguese coastal site, MAPE was significantly higher. This likely results from more complex system dynamics (i.e. land outflows, upwelling effects, etc.) and reduced availability of quality satellite data for model training (due to near-land interference). Nevertheless, focusing on a one-month prediction window, MAPE was <2% for this period indicating the applicability of the approach for shorter-term prediction.

GAIN results demonstrate the viability of applying ML-based approaches, addressing transferability, biases and robustness by combining feature selection and disparate models with specific characteristics in a weighted aggregation based on average model performance.

Results demonstrate high predictive skill with low computational cost and a framework which is easily parametrised to other geographical locations. The low computational cost of the approach has many advantages. First, it enables separation of SST forecasting models from HPC centres. The suite of models presented here can be trained on a laptop and applied to any geographical location. Once trained, the inference step is of negligible computational expense and can be readily deployed on edge-type devices (e.g. in-situ devices deployed in the ocean).

Ensemble-based forecasting is a widely used technique to account for uncertainty inherent in numerical modelling studies. Leveraging multiple simulations that encompass a wide range of potential representations and scenarios facilitates an expanded exploration of likely future conditions and provides probabilistic information on forecasts. Combining ensemble forecasting with aggregation provides a single deterministic forecast that 1) has higher predictive skill than single, best performing model and 2) reduces various model prediction to a single forecast that is actionable for stakeholders. This is typically done with some form of averaging across all ensemble members or selection of the best individual model (based on some metric). A machine learning based aggregation approach as presented here outlines a comprehensive technique that leverages information from past model performance and observations to aggregate ensemble elements into a single forecast. The non-invasive framework can be easily integrated into an on-line operational forecasting system (i.e. into

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

the IBM Bluemix service for aquaculture). This can be readily extended to other models and in particular to combining and aggregating models with different levels of complexity and different fundamental physics (e.g., combining data-driven models with mechanistic models, or deterministic approaches with stochastic). The intersection of data, machine learning and ensemble aggregation with Cloud service platform provides framework for rapid dissemination of accurate, actionable forecasts to stakeholders.

5. Conclusions

Extracting value from data in aquaculture is a combination of instrumentation, communication, curation, integration and modelling. As part of the GAIN project we aim to exploit the capabilities of sensors, Cloud, and machine learning to support ecological intensification of aquaculture. Fundamental to machine learning is data. This deliverable, together with deliverable 1.1 (Service et al., 2019) describes the data generated, curated, processed, analysed, and augmented by GAIN. We will leverage these large heterogenous datasets to develop methods and tools to better inform on dynamics within farms, and improve decision making in the aquaculture industry.

References

Bell, M. J., Schiller, A., Le Traon, P.-Y., Smith, N. R., Dombrowsky, E., & Wilmer-Becker, K. (2015). An introduction to GODAE OceanView. *Journal of Operational Oceanography*, *8*(sup1), s2–s11. https://doi.org/10.1080/1755876X.2015.1022041

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 785–794. https://doi.org/10.1145/2939672.2939785

de Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean Absolute Percentage Error for regression models. *Neurocomputing*, *192*, 38–48. https://doi.org/10.1016/J.NEUCOM.2015.12.114

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107–113. https://doi.org/10.1145/1327452.1327492

Føre, M., Frank, K., Svendsen, E., Alfredsen, J. A., Dempster, T., Eguiraun, H., … Alver, M. O. (2018). Precision fish farming: A new framework to improve production in aquaculture. *Biosystems Engineering*, *173*, 176–193. https://doi.org/10.1016/J.BIOSYSTEMSENG.2017.10.014

George, L. (2011). *HBase: the definitive guide: random access to your planet-size data*. Retrieved from

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

https://books.google.co.uk/books?hl=en&lr=&id=nUhiQxUXVpMC&oi=fnd&pg=PR7&dq=HBase:+The+Definitive+Guide&ots=Hhe5jPWpWd&sig=idJM4gptqHBBxd8vJKdE9QndC8I

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Retrieved from https://books.google.co.uk/books?hl=en&lr=&id=omivDQAAQBAJ&oi=fnd&pg=PR5&dq=goodfellow+deep+learning&ots=MMR2imlINV&sig=ygji347eXNZRvEcWQsbWlcNuNS8

Handeland, S. O., Imsland, A. K., & Stefansson, S. O. (2008). The effect of temperature and fish size on growth, feed intake, food conversion efficiency and stomach evacuation rate of Atlantic salmon post-smolts. *Aquaculture*, *283*(1–4), 36–42. https://doi.org/10.1016/J.AQUACULTURE.2008.06.042

Hastie, T. J., & Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, *1*(3), 297–318. https://doi.org/10.1201/9780203738535-7

Hey, A., Tansley, S., & Tolle, K. (2009). *The fourth paradigm: data-intensive scientific discovery*. Retrieved from https://www.fh-potsdam.de/fileadmin/user_upload/fb-informationswissenschaften/bilder/forschung/tagung/isi_2010/isi_programm/TonyHey_-__eScience_Potsdam__Mar2010____complete_.pdf

Hirahara, S., Balmaseda, M., & Boisseson, E. de. (2016). *26 Sea Surface Temperature and Sea Ice Concentration for ERA5*. Retrieved from https://www.ecmwf.int/sites/default/files/elibrary/2016/16555-sea-surface-temperature-and-sea-ice-concentration-era5.pdf

Hoerl, R. W. (1985). Ridge Analysis 25 Years Later. *The American Statistician*, *39*(3), 186–192. https://doi.org/10.1080/00031305.1985.10479425

Kilpatrick, K. A., Podestá, G., Walsh, S., Williams, E., Halliwell, V., Szczodrak, M., … Evans, R. (2015). A decade of sea surface temperature from MODIS. *Remote Sensing of Environment*, *165*, 27–41. https://doi.org/10.1016/J.RSE.2015.04.023

Klein, L. J., Marianno, F. J., Albrecht, C. M., Freitag, M., Lu, S., Hinds, N., … Hamann, H. F. (2015). PAIRS: A scalable geo-spatial data analytics platform. *2015 IEEE International Conference on Big Data (Big Data)*, 1290–1298. https://doi.org/10.1109/BigData.2015.7363884

Martin, M. J., Balmaseda, M., Bertino, L., Brasseur, P., Brassington, G., Cummings, J., … Weaver, A. T. (2015). Status and future of data assimilation in operational oceanography. *Journal of Operational Oceanography*, *8*(sup1), s28–s48. https://doi.org/10.1080/1755876X.2015.1022055

Mello, U., & Raghavan, S. (2018). Smarter Farms: Watson Decision Platform for Agriculture. Retrieved April 7, 2019, from https://www.ibm.com/blogs/research/2018/09/smarter-farms-agriculture/

Newman, S. (2015). Building Microservices DESIGNING FINE-GRAINED SYSTEMS. In *Building Microservices*. https://doi.org/10.1109/MS.2016.64

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

O'Donncha, F., Hartnett, M., Nash, S., Ren, L., & Ragnoli, E. (2015). Characterizing observed circulation patterns within a bay using HF radar and numerical model simulations. *Journal of Marine Systems*, *142*. https://doi.org/10.1016/j.jmarsys.2014.10.004

O'Donncha, F., Zhang, Y., Chen, B., & James, S. C. (2018). An integrated framework that combines machine learning and numerical models to improve wave-condition forecasts. *Journal of Marine Systems*, *186*, 29–36. https://doi.org/10.1016/J.JMARSYS.2018.05.006

Service, M., Grant, J., Icely, J. D., Lopez, M., Moore, H., Micallef, G., … Zhu, C. (2019). *Report on Instrumentation of GAIN pilot sites. Deliverable 1.1. GAIN - Green Aquaculture INtensification in Europe*.

Tarkoma, S. (2012). *Publish/subscribe systems: design and principles*. Retrieved from https://books.google.co.uk/books?hl=en&lr=&id=iLGzgqi5JPgC&oi=fnd&pg=PT14&dq=tarkoma+2012&ots=-Bl0yzTIrm&sig=Yw6EwwDhdBVjZkVlRQgUVj6nVks

Team, R. C. (2013). *R: A language and environment for statistical computing*. Retrieved from ftp://ftp.uvigo.es/CRAN/web/packages/dplR/vignettes/intro-dplR.pdf

von Schuckmann, K., Le Traon, P.-Y., Smith, N., Pascual, A., Brasseur, P., Fennel, K., … Zuo, H. (2018). Copernicus Marine Service Ocean State Report. *Journal of Operational Oceanography*, *11*(sup1), S1–S142. https://doi.org/10.1080/1755876X.2018.1489208

Wen, T., Gasic, M., Mrkšić, N., Su, P., Vandyke, D., & Young, S. (2015). Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. *2015 Conference on Empirical Methods in Natural Language Processing*, 1711–1721. Retrieved from https://www.aclweb.org/anthology/D15-1199

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82. Retrieved from http://georgemaciunas.com/wp-content/uploads/2012/07/Wolpert_NLFoptimization-1.pdf

Xingjian, S., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*, 802–810. Retrieved from http://papers.nips.cc/paper/5955-convolutional-lstm-network-a-machine-learning-approach-for-precipitation-nowcasting

Zhao, Z., Chen, W., Wu, X., Chen, P. C., & Liu, J. (2017). LSTM network: a deep learning approach for short-term traffic forecast. *IET Intelligent Transport Systems*, *11*(2), 68–75. Retrieved from https://ieeexplore.ieee.org/abstract/document/7874313/

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330

List of Figures

The project has received funding from the European Union's Horizon 2020 Framework Research and Innovation Programme under GA n. 77330