frontiers in Animal Science

ORIGINAL RESEARCH
published: 28 July 2021
doi: 10.3389/fanim.2021.695054

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]*

[1] IBM Research Europe, Dublin, Ireland, [2] Department of Oceanography, Dalhousie University, Halifax, NS, Canada, [3] Department of Oceanography, University of Stirling, Stirling, Scotland, [4] Gildeskål Research Station, Inndyr, Norway, [5] Cooke Aquaculture, Orkney, United Kingdom, [6] Marine Affairs Program, Dalhousie University, Halifax, NS, Canada

Aquaculture, or the farmed production of fish and shellfish, has grown rapidly, from supplying just 7% of fish for human consumption in 1974 to more than half in 2016. This rapid expansion has led to the growth of Precision Aquaculture concept that aims to exploit data-driven management of fish production, thereby improving the farmer's ability to monitor, control, and document biological processes in farms. Fundamental to this paradigm is monitoring of environmental and animal processes within a cage, and processing those data toward farm insight using models and analytics. This paper presents an analysis of environmental and fish behaviour datasets collected at three salmon farms in Norway, Scotland, and Canada. Information on fish behaviour were collected using hydroacoustic sensors that sampled the vertical distribution of fish in a cage at high spatial and temporal resolution, while a network of environmental sensors characterised local site conditions. We present an analysis of the hydroacoustic datasets using AutoML (or automatic machine learning) tools that enables developers with limited data science expertise to train high-quality models specific to the data at hand. We demonstrate how AutoML pipelines can be readily applied to aquaculture datasets to interrogate the data and quantify the primary features that explains data variance. Results demonstrate that variables such as temperature, wind conditions, and hour-of-day were important drivers of fish motion at all sites. Further, there were distinct differences in factors that influenced in-cage variations driven by local variables such as water depth and ambient environmental conditions (particularly dissolved oxygen). The framework offers a transferable approach to interrogate fish behaviour within farm systems, and quantify differences between sites.

Keywords: machine learning, hydroacoustic, aquaculture, AutoML, IoT

OPEN ACCESS

Edited by:
Suresh Neethirajan, Wageningen University and Research, Netherlands

Reviewed by:
Isabella Condotta, University of Illinois at Urbana-Champaign, United States
Matthew Whitney Jorgensen, Agricultural Research Service, United States Department of Agriculture, United States

****Correspondence:***
Fearghal O'Donncha
feardonn@ie.ibm.com

Specialty section:
This article was submitted to Precision Livestock Farming, a section of the journal Frontiers in Animal Science

Received: *14 April 2021*
Accepted: *01 July 2021*
Published: *28 July 2021*

Citation:
O'Donncha F, Stockwell CL, Planellas SR, Micallef G, Palmes P, Webb C, Filgueira R and Grant J (2021) Data Driven Insight Into Fish Behaviour and Their Use for Precision Aquaculture. Front. Anim. Sci. 2:695054. doi: 10.3389/fanim.2021.695054

1. INTRODUCTION

1.1. Background

Salmon fish farming started on an experimental level in the 1960s but became an industry in Norway and the UK in the 1980s, and in Chile in the 1990s (Laird, 1996). Global salmon production is currently circa 2.4 Million Tonnes per annum in 2018 (FAO, 2020) with a market value of approximately 16 billion euros (Planet Tracker, 2021). Current production is mainly concentrated in Norway, Chile, UK and Canada. The intensification of the salmon industry requires more specific

knowledge of the use of feed and the control and response to the environment. The individual number of animals used in aquaculture has increased substantially over the last 3 decades. For example, only for Scottish aquaculture the number of fish transferred to sea increased from 25 million in 1990 to 47 million in 2018 (Marine Scotland Science, 2018).

The use of big data and models to guide production in the *agriculture* space is well established, with initial implementations of *precision agriculture*, or information-based management of agricultural production systems beginning in the 1980s. The fundamental approach reduces to leveraging data from disparate sources (satellite, sensor arrays, image data, etc.) to guide decision and apply treatment in the right place at the right time. While precision agriculture originated with crop productions, applications related to livestock farming subsequently blossomed. These generally related to leveraging various sensor technologies to monitor health and productivity of livestock. Examples include radio frequency identification (RFID) tags to identify cattle in computer-controlled feeders, milking robots that ease the work of dairy operators, and automatic milk feeders that customise the milk supplement for calves, measure body weight and body temperature and generate reports (Gebbers and Adamchuk, 2010).

Precision *aquaculture* on the other hand is a nascent management concept that requires adoption of technologies from both crop and livestock management systems. Namely, an effective aquaculture management system needs to understand the environmental conditions *and* the conditions of fish within that cage. Modern fish farms are comprised of cages with up to 200,000 fish. As farms are typically composed of 10–20 cages, and multiple farms are often co-located in a bay, the total number of individual fish is enormous. This precludes the direct translation of concepts from livestock farming, and in practise, precision aquaculture is a marriage of approaches developed for both precision livestock and grain cultivation, i.e., fish are not managed as individuals as are cows, yet are obviously more complex in management than plants (O'Donncha and Grant, 2019).

There has been extensive literature on observing, modelling and quantifying the environmental conditions of aquaculture systems to inform on aspects such as site carrying capacity (Ferreira et al., 2013), environmental impacts (Buschmann et al., 2006), and mitigation activities to reduce environmental footprint (Costa-Pierce, 2003). These class of studies contribute to the first pillar of precision aquaculture; of equal importance is monitoring and extracting insight on fish behaviour to inform operations. In this paper, we investigate fish behaviour within a cage using hydroacoustic sensors. These datasets provide high resolution estimates of fish motion at the biomass (or group) level, and allow inference on fish behaviour and implications for farm management. Such data and the information they provide can be fundamental to empowering precision aquaculture and data-driven decision making.

This paper presents an analysis of environmental and hydroacoustic data from three salmon farms in Norway, Scotland, and Canada. We use statistical and machine learning approaches to interrogate the primary environmental drivers of changes in fish behaviour (i.e., variations in vertical motion and distributions). The contributions of the paper are as follows:

- We describe the monitoring, collection, and statistical analysis of environmental and hydroacoustic datasets at three salmon farms with very different environmental and geographical characteristics.
- We outline an AutoML (or automatic machine learning) model development pipeline (data curation, preprocessing, and model setup) to train and deploy a machine learning model to forecast fish behaviour using a "no-code" paradigm.
- We present a framework to interrogate the trained model and extract insight into the environmental drivers of fish behaviour using explainable machine learning techniques. We discuss the results of the data driven interrogation of fish response against known drivers of fish behaviour from literature.

The objective of the paper was to develop a transferable approach to interrogate caged salmon behaviour and inform farm management. Results indicate that data-driven approaches have great promise to provide automated insight into fish behaviour, and the environmental conditions that influence that behavioural response. However, the analysis needs to be informed by domain expertise from ecology and fish welfare to allow a robust interpretation of results.

The paper is structured as follows: below, we present a detailed overview and literature review of fish behaviour and how information on fish movement and vertical distribution provides insight into behavioural and welfare aspects. Section 2 describes the study sites, introduces the machine learning approaches used, and outlines the data curation and analysis framework. Section 3 presents results from this study while finally, we outline conclusions from this research and makes recommendations for further work.

1.2. Behaviour

Many studies of fish behaviour are in the wild or under controlled laboratory conditions. Few of them have been done under farmed conditions mainly due to the challenges and multiple restrictions this production systems poses to the researchers (Johansson et al., 2006, 2014). However, a greater understanding of the role of fish behaviour as a key health and welfare indicator is essential to allow more autonomous monitoring of fish health. The importance of fish behaviour as a farm management tool has spurred interests in new technologies to monitor and infer behaviour such as sonar and video images or the use of artificial intelligence.

1.2.1. Salmon Aquaculture

Many different biological, environmental and social parameters influence the behaviour of salmon when farmed in sea cages. Parasites, such as sea lice, are a biological example that can cause behavioural changes to farmed salmon in order to combat infestation (Bui et al., 2016). Sea lice are concentrated near the surface, and methods to limit fish surface time have been a focus of farm mitigation activities. Salmon have been observed to prefer deeper depths once highly infested to avoid further infestation (Bui et al., 2016). Temperature and dissolved oxygen

(DO) are an example of environmental parameters that affect the behaviour of Atlantic salmon. For example, Atlantic salmon will distribute themselves according to preferred temperature range, 8–18°C, with changes in behaviour occurring above and below the threshold (Johansson et al., 2006, 2009; Oppedal et al., 2011). Similarly, DO is an important parameter affecting fish behaviour as concentrations below the optimal range cause physiological stress and related behavioural changes such as a reduction in feeding (Dempster et al., 2016; Oldham et al., 2018). Light intensity is another major contributor to fish behaviour by changing vertical distribution. During daylight hours, when light intensity is at its greatest, fish tend to swim deeper in net pens to avoid surface predators (Fernö et al., 1995; Oppedal et al., 2001). It has been hypothesised ascent toward the surface during nighttime is a photo-regulatory behaviour to maintain schooling as light fades. Furthermore, seasonal variation in light availability changes vertical distribution with winter swimming depths generally shallower than summer swimming depths (Juell and Westerberg, 1993; Oppedal et al., 2001). However, this diel seasonal pattern changes when surface mounted artificial lights are installed in net pens (Oppedal et al., 2001). Populations have also been observed to remain in the upper half of net pens, under normal stocking densities, to avoid large piscivorous fish present under net pens (Fernö et al., 1995; Juell and Fosseidengen, 2004; Johansson et al., 2006, 2009; Dempster et al., 2016; Føre et al., 2017). Additionally, hydrodynamic conditions, such as waves and currents, will affect vertical distribution with stronger waves encouraging fish toward the surface (Johannesen et al., 2020).

A group of fish that voluntarily remain together, or a shoal, will have social group behaviour (Martins et al., 2011). A shoal will adopt a polarised swimming pattern in order to minimise the possibility of collisions and by synchronising these patterns a shoal can be deemed a school (Oppedal et al., 2011). Within a school there are rules in which all individuals must follow, and deviations from these rules by one or a few individuals can result in a group reaction. As with any population, individuals within a school will react differently when placed under the same stressors. For example, one fish may be more motivated by feed and swim toward the surface while another may be content waiting for feed to fall. These individual differences can affect the behaviour of the whole of the shoal regarding the responses to environmental parameters or other external or internal stressors. The internal state of the fish is also an important parameter to consider to understand the fish behavioural responses (Castanheira et al., 2017; Damsgård et al., 2019). Internal states being the final behavioural decision maker for the animal to respond to external stimulae (Huntingford et al., 2011).

1.3. Related Work

Technology on farms has increased significantly in the past decade (Føre et al., 2018), and ongoing efforts focus on the improvement of fish welfare and optimisation of farm operations, e.g., minimising the waste of feed. Sensors such as real-time oxygen and temperature probes, or acoustic tracking of fish are becoming more common in fish farming. The use of hydroacoustic sensors to infer the behavioural response of the fish to the physical structure of the cage, aquaculture practises, and the external environment can be a highly reliable and non-invasive operational welfare indicator (OWI) (Martins et al., 2011; Damsgård et al., 2019). A key area of research for the industry is whether sensor observations such as these can be used to augment welfare indices, thereby reducing the necessity to collect cumbersome manual samples such as lice counts and gill health.

Traditional methods of *in situ* observations of fish behaviour include visual inspections of the fish through random sampling, and video cameras placed in the feed zone of cages (Føre et al., 2018). Visual inspections are difficult to achieve on large populations and can be stressful to the fish (Martins et al., 2011). Video cameras can help identify when feed falls below the depth of the camera indicating the fish within the cage are satiated. Although these videos provide real-time images, they only supply a small frame of the cage and are hindered by the challenge of capturing high quality images underwater. In order to study behaviour on the cage's population, a larger view of the cage is required. One suggested method of continuous observations includes a commercial system, CageEye (2021). CageEye is a hydroacoustic sensor which is placed in the cage and captures (in real-time) the relative density of fish in the water column. Previous studies (Juell et al., 1993; Lindem and Houari, 1993) have investigated the effectiveness of using CageEye to completely automate feeding by using the detection of fish depth as indication of appetite. However, the use of this technology has potential to be used to study fish behaviour and indirectly welfare.

The association between behaviour and welfare can be determined for Atlantic salmon by understanding abnormal behaviour as it is linked with stressors (Martins et al., 2011; Damsgård et al., 2019). Therefore, continuous monitoring of fish behaviour can provide a more comprehensive perception of environmental conditions and their effect on welfare. The description, classification, and understanding of fish movement, as well as the environmental stimuli responsible for that behaviour could become the foundation for the creation of an early-warning system of fish welfare. This early-warning system can trigger changes in aquaculture practises that result in improved welfare conditions for farmed fish.

There are numerous studies dedicated to using machine learning to characterise and manage animal behaviour in *agriculture* (Liakos et al., 2018). These include automated monitoring systems based on video camera (Matthews et al., 2017), and prediction of bovine weight trajectories based on historical data (Alonso et al., 2015). Faced with a more difficult monitoring environment, applications of machine learning to aquaculture have developed slower. Many studies have investigated how machine learning could improve ocean monitoring and forecasting either by mining large ocean datasets Gokaraju et al. (2011) or relating future conditions to historical observations (Wolff et al., 2020). More recently, the applications of machine learning and computer vision technology to aquaculture is receiving a lot of attention. Broadly these are applied across two categories: 1) pre-harvest and during cultivation, and 2) post-harvest (Saberioon et al., 2017). In the pre-harvest and cultivation stage, much of the research

focuses on monitoring fish behaviour. A number of studies have demonstrated accurate monitoring of fish behaviour and trajectory (Kato et al., 2004; Pérez-Escudero et al., 2014), although these have been predominantly laboratory-based. Monitoring and optimising feeding activities using computer vision and machine learning is an active area of research (Atoum et al., 2014). However, these provide little information about behavioural dynamics during feeding and are still at an early stage of development (Oppedal et al., 2011; Saberioon et al., 2017). Saberioon et al. (2017) provides an excellent review of applications of computer vision and machine learning technologies to aquaculture.

2. MATERIALS AND METHODS

Hydroacoustic methods provide a proxy measure for density and distribution of marine animals in form of acoustic backscattering (Foote, 2009). The fundamental principle is based on emitting a signal of known type and power level from a transducer. As it encounters regions of the medium with differing properties, also called heterogeneities, the sound is generally redistributed, or scattered, in all directions. This makes possible detection of the scattered sound with transducer and suitable receiver electronics. Advantages linked to hydroacoustic sampling techniques include, high spatial and temporal resolution, autonomous long-term sampling duration, range (especially during poor visibility when visual-based methods tend to fail), and a non-invasive surveying approach (Scherelis et al., 2020). Given these advantages, hydroacoustics is increasingly used to characterise animal behaviour in the marine environment, and considered a promising system to improve management of aquaculture farms (Bjordal et al., 1993; Juell et al., 1993).

In this study, hydroacoustic data were collected by one of two sensors "CageEye" (Scherelis et al., 2020) or "Aquaculture Biomass Monitor" ABM (2020). Broadly speaking, processed hydroacoustic data generates two metrics: volume backscattering strength (S_v), is often considered as a proxy for fish biomass; while target strength (TS) is an acoustic measure of fish length (Simmonds and MacLennan, 2008). TS is a measure of the acoustic reflectivity of a fish, which varies depending on the presence of a swim bladder and on the size, behaviour, morphology, and physiology of the fish. These outputs can be used to generate estimates of fish density and biomass (Boswell et al., 2007) within a cage.

2.1. Study Sites

This study considers three salmon cage farms in Norway (NOR), Scotland (SCO), and Canada (CAN). For each site a number of environmental sensors were deployed monitoring a range of parameters, including temperature, DO, and current speed. These were complemented with weather data from *in-situ* weather stations or model generated reanalysis from IBM Environmental Intelligence Suite available through their public API IBM (2021b).

2.1.1. Norway Site

The Norwegian site at Røssøya Nord (coordinates: 67° 4.38′ N 13° 56.855′ E) is a commercial farming site owned by GIFAS in Gildeskl municipality (Nordland, Norway). The site has a mooring system with placements for 16 cages (cage circumference 90 m; maximum cage depth 27 m). Seven cages (circumference 90m, max depth 20 m) were stocked with 18G S1 smolt produced at Salten Smolt AS and Helgeland Smolt AS in August/September 2018. Smolt were transferred at an average weight of 61 g (Salten) and 122 g (Helgeland) and each cage was stocked with approximately 150,000 smolt. The fish density at stocking was 1.3–2.1 kg/m^3.

All the cages were fed on a standard commercial diet (EWOS Robust). Feed was delivered from the silos on the barge to the cages via an air-blowing system and into a rotor spreader at the cage surface. Feeding start time, feeding intensity, and number of meals were adjusted according to day-length, weather, and observations on fish behaviour by the site staff.

A designated cage was instrumented to explore environmental variations and fish distribution (hydroacoustic sampling) between 20/02/2019 and 800 is a numerical ocean 31/10/2019. During the experiment it was decided to change the fish stock since they were not growing optimally. Therefore, the stock were replaced with 50,000 fish supplied by Helgeland Smolt in July 2019. The fish density after the change in fish stock was 10.1 kg/m^3.

Environmental variables were measured by Aanderaa instruments (Xylem analytics, Norway). The variables measured were salinity, temperature, DO, as well as current speed and direction. Animal variables were monitored by an Aquaculture Biomass Monitor sensor, ABM (Biometrics AS, Norway) which consists of a split-beam sonar mounted on a buoy and it can detect over 50,000 fish per day. The sonar provided an estimate of total biomass, biomass distribution with depth, fish weight distribution and fish swimming speed. Fish position and distribution are reported hourly, while estimates of average fish size are returned at multi-day period. **Figure 1** presents a schematic of typical sensor configuration within a cage. The hydroacoustic sensor was installed beneath the cage looking upwards. It sampled at an angle of approximately 42° from horizontal. An environmental sensor was deployed at approximately mid-depth in the cage.

Since the environmental sensor deployment (May–July) did not cover the entire study period, we augmented environmental data with output from the NorKyst-800 ocean model (Albretsen et al., 2011). The NorKyst-800 is a numerical ocean modelling system deployed to simulate physical oceanography variables such as sea level, temperature, salinity and, currents for all coastal areas in Norway and adjacent seas. The model has a horizontal resolution of 800 m, 35 layers in the vertical, and can be downloaded from the OPeNDAP server provisioned by the Norwegian Meteorological Institute (Norwegian Meteorological Institute, 2021). The model provides a satisfactory representation of Norwegian off- and onshore dynamics but requires higher resolution to resolve the dynamics of most Norwegian fjords properly (Albretsen et al., 2011).

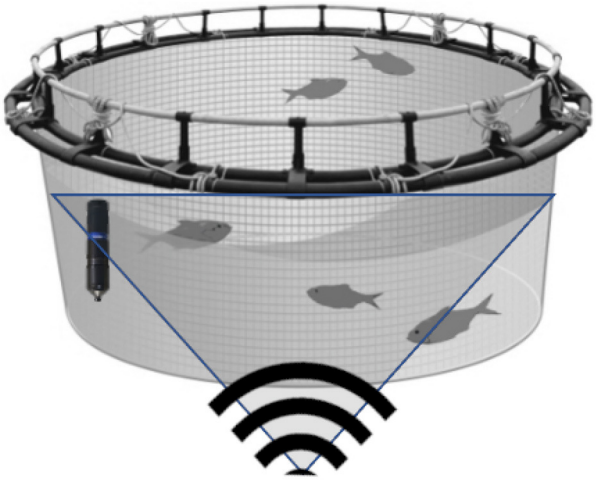

FIGURE 1 | Schematic of sensor configuration within a cage. The environmental sensor is denoted in the left hand side of cage with the exact location varying across sites (e.g., for Norway it was deployed at 5 m depth in centre of cage), while the hydroacoustic sensor and it's sampling domain is denoted beneath the cage (pointing upwards). Note, that the hydroacoustic sensor contains sampling "blind spots," especially toward the bottom of the cage which causes challenges in shallow sites where depth is limited. This figure represents an idealised representation of the deployment and naturally there are many practical considerations with regards sensor deployment (power source, ease of access, anchoring point, etc.).

2.1.2. Scotland Site

The salmon sea farm is located at Carness Bay, Orkney (coordinates: 59° 00.637′ N 02° 55.374′ W). This barge fed site includes a total of 12 × 100 m circumference cages and maximum net depth of 6 m. The total number of fish stocked at the end of January 2019 was approximately 230,000, with a stocking density of 5.1 kg/m^3. Fish stocked were 18S1 smolts, which were moved from Meil bay to Carness Bay on 5th February 2019. Each cage is stocked with approximately 20,000 salmon.

All the cages were fed on a standard commercial diet (Biomar, power extreme). Feed was distributed to the fish cage from the hulls on the barge via an air-blowing system. Feeding was controlled by AkvaConnect software (AkvaGroup AS). Feeding started at pre-determined times according to day length, most often with a meal duration of 30–60 min. The stock were monitored by camera and feeding intensity was adjusted accordingly. For example, should the fish exhibit poor feeding activity, the feeding intensity would be decreased or stopped in the affected cages. Adjustments in the feeding intensity were done daily according to requirement and light availability. As soon as there was enough light, the first meal started (lasting 30 min according to appetite), and fish were fed again after a pause of 4–6 h. A maximum of 2 meals were administered per day.

Cage 8 was instrumented to collect both environmental and animal variables i.e., variables concerning fish behaviour, growth and welfare data. Water temperature, DO, turbidity and salinity were measured for each cage daily by the site management staff. Cage 8 was monitored with two sensors to record DO and temperature: a handheld underwater wireless sensor (OxyGuard International, 2014) sampled daily from 08/01/2019 to 30/10/2019 and an *in-situ* Realtime Aquaculture sensor (Innovasea, 2021) sampled at 10-min intervals at a depth of 2.5 m from 13/11/2019 to 07/02/2020. As for the Norway site, due to the difficulty to collect continuous environmental data for the entire period, model data extracted from the Copernicus Marine Service model repository augmented sensor datasets. Temperature, DO, sea surface height (SSH), and current speed were extracted from the Atlantic North West Shelf model at the surface layer. Data is available at a 1.5 km (Tonani et al., 2019) horizontal resolution at hourly intervals and can be freely downloaded from the Copernicus portal. We compared sensor and model data for the period and since the overall trends were very similar, we augmented periods when sensor data were missing with Copernicus data. We noted that the model tended to overestimate magnitude of temperature, but since they captured the temporal variations, it served to adequately represent conditions for machine learning purposes. It's worth noting that the ML model is not affected by data magnitudes since data is normalised. Instead it learns how the predictand (label) varies in response to predictors (features). Using data from a physics model can be a pragmatic approach to handling missing values in environmental studies, thereby avoiding reliance on statistical imputation methods.

The relative biomass distribution within the cage was assessed using the beam sonar system, CageEye. The system in Orkney was made up of an echosounder and one transducer. The transducer was placed in the cage at a depth of approximately 5.5 m, and connected to the echosounder cabinet, which was placed on the cage ring and sent the data wirelessly to the base station at the feeding barge. The transducer was placed as deep as possible, looking up at most of the biomass of the cage. The transducer had two angles that they switch between, approximately 14 degrees (200 kHz) and 42 (50 kHz) degrees: this allowed one to get echogram recordings of both. It is important to note that this site did not have power at night, and consequently (since sensor did not have battery source), the acoustic sensor was switched off between 18:00 and 06:00.

2.1.3. Canada Site

The Canadian study site located in Saddle Island, Nova Scotia (coordinates: 44° 30.225′ N 64° 2.923′ W) is a commercially operated Atlantic salmon farm. The site had one column of 6 cages measuring 150 m circumference and a maximum depth of 11 m. Each cage contained approximately 60,000 fish with a stocking density of about 10 kg/m^3. Fish were fed twice daily, with the exact times dependent on daylight availability.

Each cage was equipped with two RealTime Aquaculture (Innovasea, 2021) probes deployed at 2 and 8 m depths. The probes measured temperature and DO, while an ADCP profiler sampling current speed was deployed in the northeast corner of the farm. Sea surface height was extracted from the Copernicus portal, in similar manner to the other two sites. Two of the cages were equipped with a CageEye sensor from 11/09/2019 to 30/10/2019. Each system consisted of three transducers, with two placed in opposite corners at 7 m depth, facing upwards,

TABLE 1 | Summary metrics for the three sites describing location, water depth, cage depth, average tidal range, number of fish per cage, and fish density.

	NOR	SCO	CAN
Latitude	67° 4.38′ N	59° 00.637′ N	44° 30.225′ N
Longitude	13° 56.855′ E	02° 55.374′ W	64° 2.923′ W
Water depth (m)	60	10	11
Cage depth (m)	27	6	11
Tidal range (m)	2.5	1.75	1.3
Fish per cage (-)	50–150,000	20,000	60,000
Fish density (kg/m^3)	1.3–10.1	5.1	10

Note that the fish stock were changed at the NOR site, hence we provide the range of values over the period.

and one near the surface, facing downwards. **Table 1** presents summary metrics on the three sites considered, while **Table 2** describes the data collected at sites. Data can be categorised along hydroacoustic and environmental sensor measured variables, and model or product data from ocean or weather model.

2.2. Machine Learning

Given sufficient data, machine learning (ML) models have the potential to successfully detect, quantify, and predict various phenomena in the geosciences. While physics-based modelling involves providing a set of inputs to a model which generates the corresponding outputs based on a non-linear mapping encoded from a set of governing equations, supervised machine learning (ML) instead learns the requisite mapping by being shown large number of corresponding inputs and outputs. In ML parlance, the model is trained by being shown a set of inputs (called features), and corresponding outputs (termed labels), from which it learns the prediction task—or in our case, we wish to predict the distribution of fish in a cage (as sampled by hydroacoustic sensor) based on a set of environmental measurements or features.

Classical works in machine learning and optimisation, introduced the "no free lunch" theorem Wolpert and Macready (1997), demonstrating that no single machine learning algorithm can be universally better than any other in all domains—variance tradeoff in effect, one must try multiple models and find one that works best for a particular problem. Selection of the most suitable algorithm and algorithmic settings is one of the most complex aspects of machine learning applications and highly dependent on user skill. An alternative approach leverages advanced *automatic machine learning* (AutoML) frameworks that aims to *learn how to learn* (Drori et al., 2018). AutoML systems uses a variety of techniques, such as differentiable programming, tree search, evolutionary algorithms, and Bayesian optimisation, to find the best machine learning pipelines for a given task and dataset (Drori et al., 2018). In this paper we applied IBM AutoAI (IBM, 2021a) to the data collected at the aquaculture sites. IBM AutoAI is a technology directed at automating the end-to-end AI Lifecycle, from data cleaning, to algorithm selection, and to model deployment and monitoring in the ML workflow (Wang et al., 2020).

As a benchmark, we compared results against a manually tuned machine learning model, namely Random Forest (RF). RF is one of the most popular machine learning models and has demonstrated excellent performance in complex prediction problems characterised by a large number of explanatory variables and nonlinear dynamics. RF is a classification and regression method based on the *aggregation* of a large number of decision trees. Decision trees are a conceptually simple yet powerful prediction tool that breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The resulting intuitive pathway from explanatory variables to outcome serves to provide an easily interpretable model.

In RF Breiman (2001), each tree is a standard Classification or Regression Tree (CART) that uses what is termed node "impurity" as a splitting criterion and selects the splitting predictor from a randomly selected subset of predictors (the subset is different at each split). Each node in the regression tree corresponds to the average of the response within the subdomains of the features corresponding to that node. The node impurity gives a measure of how badly the observations at a given node fit the model. In regression trees this is typically measured by the residual sum of squares within that node. Each tree is constructed from a bootstrap sample drawn with replacement from the original data set, and the predictions of all trees are finally aggregated through majority voting (Boulesteix et al., 2012).

While RF is popular for its relatively good performance with little hyperparameter tuning (i.e., works well with the default values specified in the software library), as with all machine learning models it is necessary to consider the bias-variance tradeoff—the balance between a model that tracks the training data perfectly but does not generalise to new data and a model that is biassed or incapable of learning the training data characteristics. Some of the hyperparameters to tune include number of trees, maximum depth of each tree, number of features to consider when looking for the best split, and splitting criteria (Probst et al., 2019).

2.3. Model Setup and Training

Data preprocessing focused on creating a curated matrix of environmental and hydroacoustic datasets to allow statistical and machine learning interrogation of relationships. Important points to consider included outlier removal, time-averaging, imputation, data augmentation, and representation of temporal dependencies). **Figure 2** summarises the data processing workflow. The hydroacoustic sensor returns estimates of fish depth at sub-second frequency. This data point reports the location (relative to the sensor) of an individual (random) fish in the cage and is based on sensor detected change in medium (water vs. flesh). For a 6-month study, these generated about 45 GB of data. Data were first grouped into 1 m bins to represent the frequency of returns at different depth levels based on the *Echo Range (m)* measurement (i.e., number of individual fish in each 1 m bin). Measurements that are outside the extents of the cage were removed as outliers, and the remaining data were then time-averaged into hourly intervals. The binned data

TABLE 2 | Synopsis of data collection at the three sites summarising the environmental variables collected and the sampling periods, source of ocean model data (used to augment sensor data), and weather data source and variables.

	NOR	SCO	CAN
Environmental sensor	Aanederaa	Realtime Aquaculture	Realtime Aquaculture
Deployment dates	21/05/2019–02/10/2019	13/11/2019–04/02/2020	16/09/2019–16/11/2019
Variables	Temperature, DO, salinity, current speed	Temperature, DO	Temperature, DO, current speed
Hydroacoustic sensor	Aquaculture Biomass Monitor	CageEye	CageEye
Deployment dates	20/02/2019–31/10/2019	01/06/2019–29/09/2019	11/09/2019–30/10/2019
Ocean model data source	NorKyst-800	Copernicus Atlantic North West Shelf	Copernicus Global Ocean
Weather data source	IBM Environmental Intelligence Suite		
Weather variables	Wind speed, air temperature, solar radiation		

were depth-averaged to generate a time series vector that is amenable toward machine learning analysis. Equation 1 was used to compute the mean of grouped data.

$$\bar{x} = \frac{\sum fx}{\sum f} \tag{1}$$

where x refers to the midpoint of depth intervals and f denotes the frequency of fish in a given interval.

Data gaps or missing values were either imputed or removed: if the gap was less than 4 h, data were imputed using a nearest neighbour linear interpolation, while if gaps were greater than (or equal) 4 h, this portion was removed from analysis (i.e., both the environmental and hydroacoustic data were removed). Autoregressive features (i.e., values at previous points in time) are often informative for machine learning models. We generated these features using 3 h sliding window size (i.e., values at previous 1, 2, and 3 h). The resulting matrix is combined with environmental data, and time-aligned. We used our open-source packages, TSML (Palmes et al., 2020) and AutoMLPipeline (Palmes, 2020) for this preprocessing step. The code we used along with the data from the NOR site is available on Github at (O'Donncha and Palmes, 2021).

As part of the machine learning model setup, we investigated two configurations:

- All features: all available environmental variables together with sliding window values of fish location data were provided to the model. Each row represents the autoregressive features together with the date-time features (year, month, day hour, day of week, etc.), and environmental data. These features are time-aligned with the corresponding label (i.e., fish location) at the desired prediction window. We setup the problem as a 1-h ahead prediction using 3 h sliding window size (i.e., autoregressive at previous 3 h were included) with 1 h stride. Due to the lack of nighttime observations we did not implement this analysis at SCO site since the incomplete daily data can reduce the insight from autoregressive analysis. Instead, at SCO site, we only considered the configuration below.
- Selected features: to interrogate the strength of relationship or dependency between fish location and environmental conditions, a subset of features were provided to the model. The subset of environmental data were selected based on analysis of literature and the features identified as being important in the first experiment above. This configuration did not include sliding window values from previous timesteps which simplified the setup (missing data no longer had to be interpolated, instead those rows could be simply dropped). It also provides more flexibility when prescribing the prediction window since once could forecast whenever environmental data is available [e.g., one could leverage the Copernicus 10-day ahead ocean forecast data (Tonani et al., 2019) to make corresponding 10-day ahead predictions]. The selected features for this configuration were: temperature, DO, current speed and direction, wind speed and direction, sea surface height, and hour of day (described in **Table 2**).

The features (environmental data primarily) and label (fish location) data were split into two groups, to form the training-data set composed of 90% of the data, and the test-data set the remaining 10%. After preprocessing and hourly-averaging, the total number of data points available were 5,847, 1,574, and 840 data points for the NOR, SCO, and CAN study sites, respectively. The data were provided to the IBM AutoAI tool (IBM, 2021a) which automatically selected the optimal combination of algorithm, feature transformations, and calibration parameters (or hyperparameters) that minimised prediction error. Mean-squared-error (MSE) was selected as the loss function to optimise. We then used the trained machine learning model to interrogate how environmental data contributed to variations in fish location and behaviour. This can be considered the true goal of the machine learning implementation, and an accurate model simply served as a means to achieve this goal.

3. RESULTS

We collected data on the observed vertical distribution of relative intensity of fish biomass within a cage at three sites. The sites were geographically disparate and had distinct characteristics in terms of both the local environment, and the farm itself that influenced fish behaviour.

Figure 3 presents summary statistics for the NOR site: the top figure shows the centrepoint of the fish biomass over the duration of the study period, while bottom figure presents a box

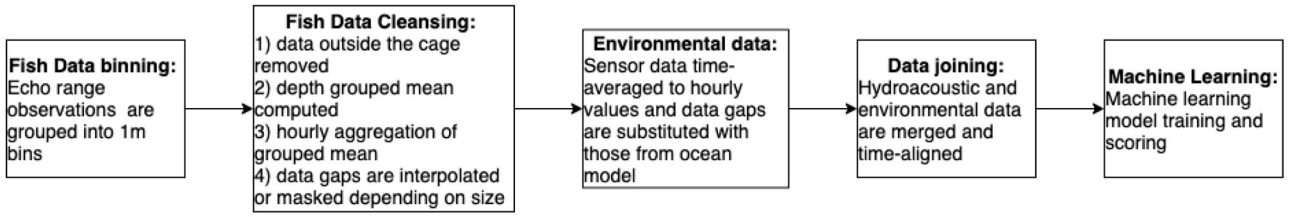

FIGURE 2 | Data processing workflow outlining the main steps of hydroacoustic data binning, data cleansing, environmental data processing, data merging, and finally machine learning model implementation.

plot elucidating hourly distribution every month. While each site exhibit unique characteristics, a number of common qualitative trends were shared, including:

- Each site demonstrated diurnal patterns to varying degrees (however the prominence of these varied across sites and at different times of year).
- Observations suggested a weak seasonal-scale pattern, with fish being at a higher position in the cage during summer months.
- A very pronounced preference for the upper portion of the cage that was independent of absolute depth. Generally fish tended to cluster in the upper one-third of the cage which can have significant implications for the density of fish in a cage (if the volume of the cage actually used by the fish is far less than the volume available).

These observations are interrogated in more detail in remainder of paper.

Figure 3 presents data from the Norwegian farm highlighting a number of noticeable trends. Firstly, despite the deep waters (approximately 60 m), and the large depth of the cage (27 m), fish tended to congregate in the upper third, and spent most of the time at depths of 3–9 m. The box plot does not show any pronounced daily pattern. It is worth noting that the northerly latitude of the site (67°N) means it is characterised by 24 h sunlight for most of the summer months which likely impeded the development of daily patterns during some of the period. Further, fish in the cage were changed between 26 and 29th July and new stock introduced, which naturally modifies recurrent patterns of behaviour. This may be the source of the more widely dispersed patterns of position evident in August, since the fish were newly introduced to the cage and conceptually displayed more chaotic behaviour patterns. Moving past these extenuating circumstances, the behaviours in September and October are possibly most indicative of typical cage-fish behaviour. These months are characterised by a weak diurnal pattern and fish congregating toward an ambient depth of about 9 m (or a third of the depth).

Figure 4 summarises information on fish distribution at the Scotland site. Both cage and water depth were significantly shallower at this farm, being 6 and 10 m, respectively. Naturally this affects the range that fish could travel and we see a quite tight clustering of average fish position between 1 and 2 m. Box plot indicates that fish sat within a tight half-metre cluster most of the time, with the box plot whiskers rarely extending outside this range. It's important to note that due to the CageEye transducer being placed inside the cage (because of the shallow water depth), some portion of fish in the cage will not be captured by the sensor. Hence, the degree of clustering is likely overestimated in this case. Results illustrate that average fish position in the cage tended to move closer to the surface during the summer period, likely influenced by warming surface temperatures. **Figure 5** includes temperature data reported at the Scotland site, illustrating warmer waters that peaked in early August before returning to moderate temperatures in September. The general trend of monthly variations in fish position, seem to follow these patterns, with July and September reporting comparable values for both temperatures and average fish positions. There was no clear diurnal pattern obvious in the data. It is important to note that due to lack of power during the night, data were not collected between the hours of 18:00 and 06:00. This naturally reduced the contribution of hour-of-day toward explaining the data.

Finally, **Figure 6** presents data from one of the cages at the CAN site. The CageEye sensor was deployed between 11/09/2019 and 30/10/2019, covering a period of large drop in temperature and reduction in daylight hours. A strong diurnal pattern was evident at this site with fish tending deeper in the cage during daylight hours. Due to the time of year the water column was not thermally stratified which may reduce the effects of temperature. While the cage depth is 10 m, the box plot illustrates that fish were generally clustered within a 2 m range and this cluster rarely goes deeper than 4 m in the cage, reflecting similar patterns to the other two sites.

Prior to more detailed statistical analysis of the data, one desires insight into the primary drivers that explains the observations. As discussed in section 2.3, machine learning models such as Random Forest provides a robust approach to efficiently explore multiple variables and associated response. We considered an analysis of the CageEye/ABM vertical distribution data from the three sites using IBM AutoAI (IBM, 2021a), automated machine learning tool. The data were preprocessed as described in section 2.3 and uploaded to the AutoAI website. The hydroacoustic data column was specified as training *labels*, and *features* were selected based on the particular experimental configuration (either "all features" or "selected features") using the AutoAI Graphical User Interface (GUI).

Our first experimental configuration ("all features" described in section 2.3) provided a wide range of environmental

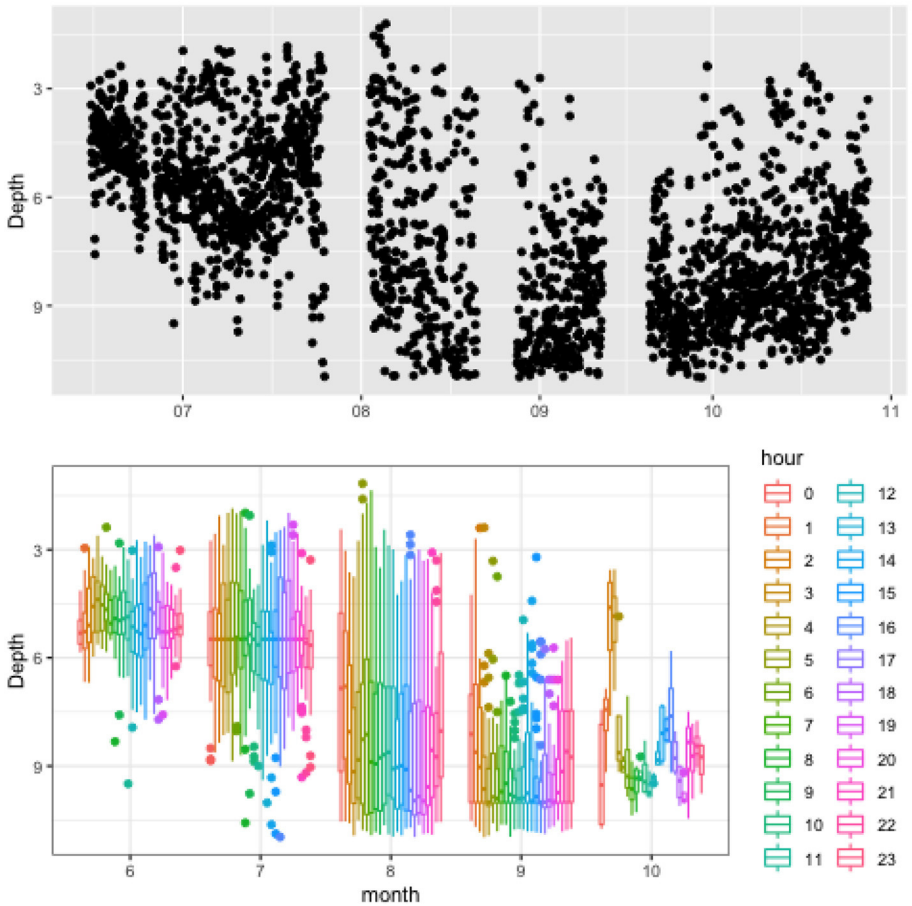

FIGURE 3 | Vertical distribution of fish in a cage at NOR farm illustrating the evolution of the centrepoint of fish biomass over the duration of study period (top) and boxplot of the data grouped into hourly intervals for each month (bottom). The box plot provides insight into distinct patterns developing in the data at different times with the color legend representing hour of day. Lines extending from the boxplot represent the range of data (i.e., minimum and maximum values), while the box section reports 25th percentile, median, and 75th percentile values. Filled circles represent outliers for the data.

(temperature, DO, current speed and direction, wind speed and direction, and sea surface height), temporal (hour of day, day of year), and autoregressive (measured fish position at 1, 2, and 3 h previously) variables as input (or features) to the model. The model was trained to make a 1-h-ahead prediction. Since the model implicitly learned to predict by learning the relationship between features and labels, we could then use the trained model to extract insight on how these features contributed to the model prediction. The resultant model demonstrate strong predictive skill reporting explained variance of 76, 81, and 75% for the NOR, SCO, and CAN sites, respectively. The relatively high correlation scores (equalling 0.87, 0.9, and 0.87, respectively) support the viability of using the model to explore the contribution that individual variables or features make toward prediction.

Figure 7 presents the *feature importance* of the supplied data to the response variable or model prediction at the CAN site (extracted from the AutoAI GUI). The feature importance measure computes the contribution or importance of each feature by calculating the increase of the model's prediction error after permuting the feature. A feature is "important" if permuting its values increases the model error, because the model relied on the feature for the prediction. A feature is "unimportant" if permuting its values keeps the model error unchanged, because the model ignored the feature for the prediction (Breiman, 2001).

Data from the CAN site provided some useful insight into salmon cage dynamics. As might be expected, autoregressive variables were a primary driver of fish behaviour. The most important feature is value at the previous timestamp ($x1$, denoting fish position 1 h previously) with $x2$ and $x3$ also contributing. Hour-of-day was the second most important feature which suggests that there was some diurnal pattern to the data that can be explained by this repeating feature. This information can serve to guide optimal feature selection for model development. Combined with domain knowledge on primary variables that influence fish behaviour (summarised in section 1.2), this information can lead to development of a more effective model. Selecting the most appropriate set of features is critical to maximising model performance, while from

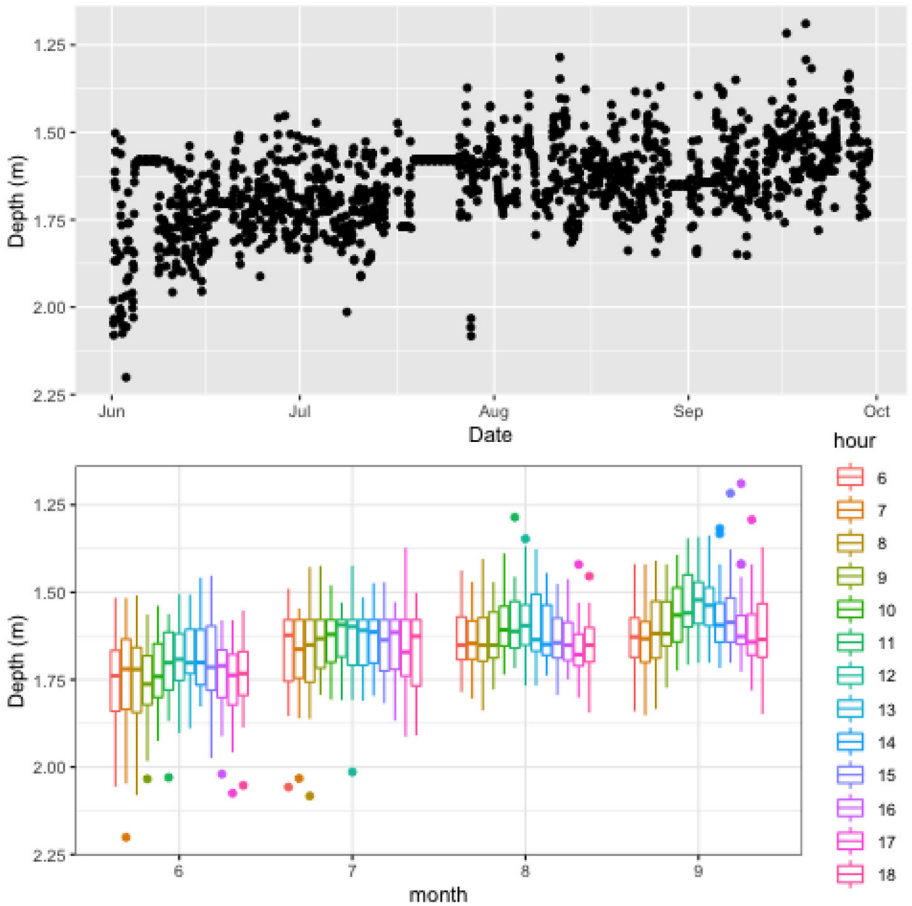

FIGURE 4 | Vertical distribution of fish in a cage at SCO farm illustrating the evolution of the centrepoint of fish biomass over the duration of study period (top) and boxplot of the data grouped into hourly intervals for each month (bottom). The box plot provides insight into distinct patterns developing in the data at different times with the color legend representing hour of day.Lines extending from the boxplot represent the range of data (i.e., minimum and maximum values), while the box section reports 25th percentile, median, and 75th percentile values. Filled circles represent outliers for the data. Note that the plots only include data between 06:00 and 18:00.

a practical point of view a model with less predictors may be more interpretable (Kuhn and Johnson, 2013).

Our second experimental setup involved a reduced set of features, namely: temperature, DO, current speed, wind speed, and salinity, together with hour-of-day. Choice of features were based on both feature importance reported in **Figure 7** and those suggested by literature. Naturally, the variance explained (or predictive skill) of the model dropped with the reduced feature set, but the analysis of feature importance or contributions can be more meaningful. The resultant model explained 59%, 64%, and 61% of variance for the NOR, SCO, and CAN sites, respectively, which represents a drop of 14–17% compared to the model with all features provided. This drop in predictive skill was balanced by an improvement in model *interpretability* and increased focus on pertinent variables (environmental conditions).

Figure 8 summarises model performance at the Canada site. It illustrates that the model captures data trends quite well reporting correlation score of 0.78. Visually, the model captures observed fish depth quite well considering the highly dynamic nature of the signal. In particular, trends in the data are adequately tracked and the model accurately replicates whether the fish move up or down in the cage in response to the provided model inputs. From a feature analysis perspective, this allows us to confidently interrogate results since we are primarily interested in variations in output rather than magnitude (i.e., changes of fish position in response to changes in environmental conditions rather than the magnitude of those changes) We used the model to understand variance explained by these drivers together with the feature importance of each. **Figure 9** presents the variable importance computed for the three locations in Norway, Scotland, and Canada.

While there were similarities in the drivers that influenced fish position at the three sites, pronounced variations existed based on the different geography and characteristics of each site. As suggested by both feature importance analysis and boxplot visualisation, time-of-day was a primary driver, particularly at the Canadian farm. This reflected the pronounced diurnal patterns that are visually evident in **Figures 3–6**, with the fish being

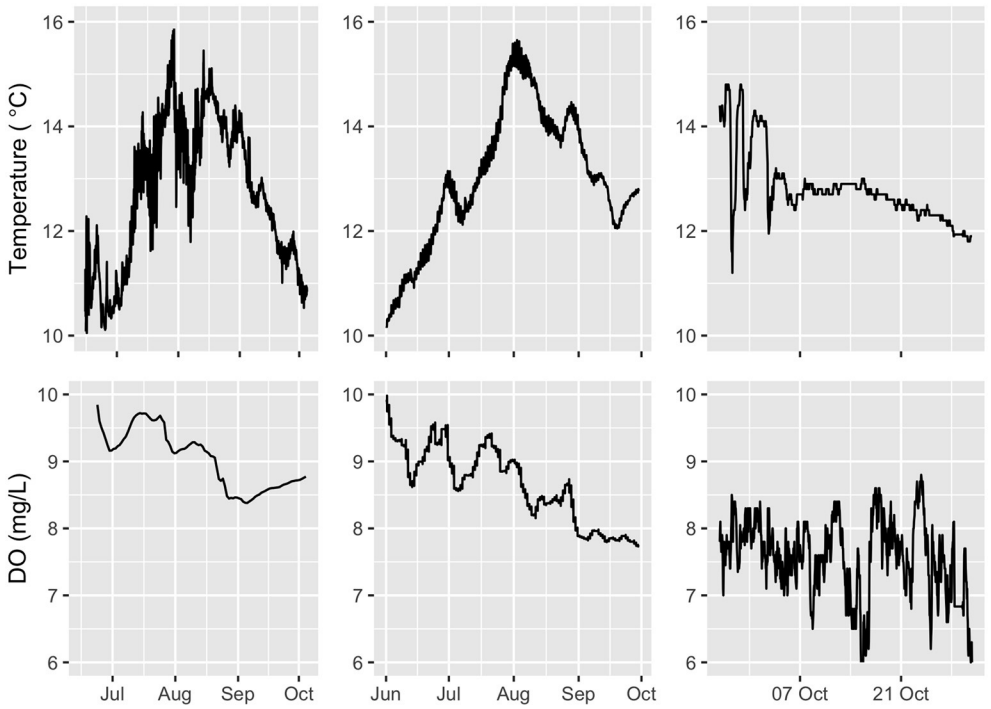

FIGURE 5 | Time series plot of temperature (top row), and DO (bottom row) reported at the three sites, NOR (left), SCO (middle), and CAN (right), respectively.

deeper in cage during daylight hours. It's worth noting that diurnal patterns were likely under represented in NOR and SCO data due to long summertime daylight hours and lack of nighttime observations, respectively. **Figure 10** presents a density plot of daytime and nighttime fish positions for both CAN and NOR (due to lack of nighttime observations, SCO was excluded). To remove the effects of long sunshine hours during June and July in NOR these 2 months were excluded from the plot. Results demonstrated a clear difference between daytime and nighttime behaviour for the CAN site and a similar but much less pronounced difference for the NOR site. In Canada, fish congregated at about 3.6 m depth and the spread around this was quite narrow during the day, while at night, fish were distributed more widely across the water column with a mean depth of 2.8 m. Similar trends were observed in Norway (although not as pronounced). The mean difference between daytime and nighttime positions were 0.52 m while fish were also more uniformly spread across the water column at night.

At all sites, physical oceanographic variables represented an important driver. Physical mixing by current speeds and wind forcing were particularly critical at the CAN site and three of the five most important variables represented physical stresses and mechanical mixing, namely current direction, wind direction, and wind speed, respectively (in order of influence). Wind stress did not represent an important driver of fish depth variance at the NOR site. This is likely due to the increased depth of cage and fish position serving to shelter from local surface dynamics. Interestingly, salinity was the primary driver of fish position at the NOR site which illustrates both fish sensitivities and local bay characteristics.

Figure 11 presents vertical profile of temperature and salinity at the site over the duration of the study period. Results illustrate a pronounced thermal stratification during the summer months, that breaks down into a well-mixed water column in spring and autumn. Variations of vertical salinity are more complex illustrating relatively low surface salinity values in September, which may be influenced by precipitation or freshwater runoff. Literature indicates that Atlantic salmon are influenced by salinity variations when younger than 3 months and during spawning periods, while indifferent to salinity at other times (Oppedal et al., 2011). The behavioural influence detected in this study may be a result of salmon expressing preference for lower salinity waters in spring, during the return migration period of salmon toward freshwater. However, **Figure 11** indicates that the vertical variation in salinity was relatively small, and additional study is necessary to understand the influence this may have on salmon variations.

While **Figures 7**, **9** provide insight into which features were important, we were interested in how the features influence the predicted outcome. A powerful approach to interrogate the variations of predictand in response to predictors are *accumulated local effects* (ALE) (Apley and Zhu, 2020). ALE quantifies the contributions of different predictors by considering the conditional probability or likelihood of changes to prediction. It has noted advantages in cases where multiple predictors are correlated and the effects are difficult to separate (which is naturally the case in ocean systems).

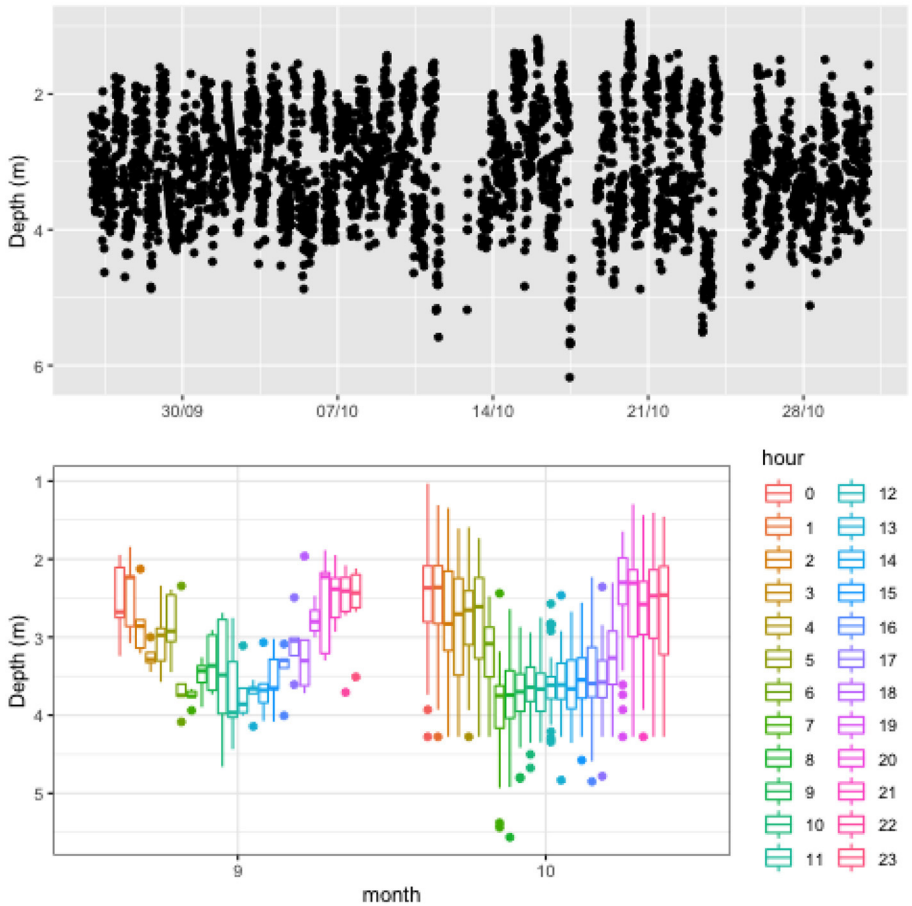

FIGURE 6 | Vertical distribution of fish in a cage at CAN farm illustrating the evolution of the centrepoint of fish biomass over the duration of study period **(top)** and boxplot of the data grouped into hourly intervals for each month **(bottom)**. The box plot provides insight into distinct patterns developing in the data at different times with the colour legend representing hour of day. Lines extending from the boxplot represent the range of data (i.e., minimum and maximum values), while the box section reports 25th percentile, median, and 75th percentile values. Filled circles represent outliers for the data.

Figure 12 presents the computed ALE for the CAN site for four variables, namely, temperature, DO, wind, and current speed to the response variable. ALE provides a quantitative way to show how the prediction (fish position) changes locally, when the feature (environmental variable) is varied. The marks on the x-axis indicates the distribution of the particular feature, showing how relevant a region is for interpretation (little or no points mean that we should not over-interpret this region). **Figure 12** allows for extraction of a number of pertinent observations on the data. The feature effects of temperature and oxygen suggest that "ambient" conditions had low importance (tends toward zero), while higher and lower values tends to trigger a response. Specifically, DO reports low importance when values were between 7–8 mgL^{-1}, while values outside this range invoke a large response by the model. It is worth noting that this large response by the fish is likely indicative of high-stress conditions. **Figure 5** plots time series of DO to illustrate the evolution at the site and localised periods when values dropped below 7 mgL^{-1}.

The contribution of wind and current speed to fish response were quite similar (as might be expected). Generally increased current speed invoked an increase in the model predicted value (i.e., fish were deeper in the cage). The plot suggests a linear relationship but is likely not enough data to draw confident conclusions on the exact relationship. This is amplified by the fact that the marks on the x-axis are quite sparse for higher values of wind and current speed indicating low number of observations for these conditions.

4. DISCUSSION AND CONCLUSIONS

The precision aquaculture concept aims to exploit data-driven management of fish production, thereby improving the farmer's ability to monitor, control and document biological processes in fish farms. The fundamental approach has been summarised as a series of steps, namely observe, interpret, decide, and act (Føre et al., 2018), that strives toward optimised operations of farms. Where precision aquaculture differs most prominently from its

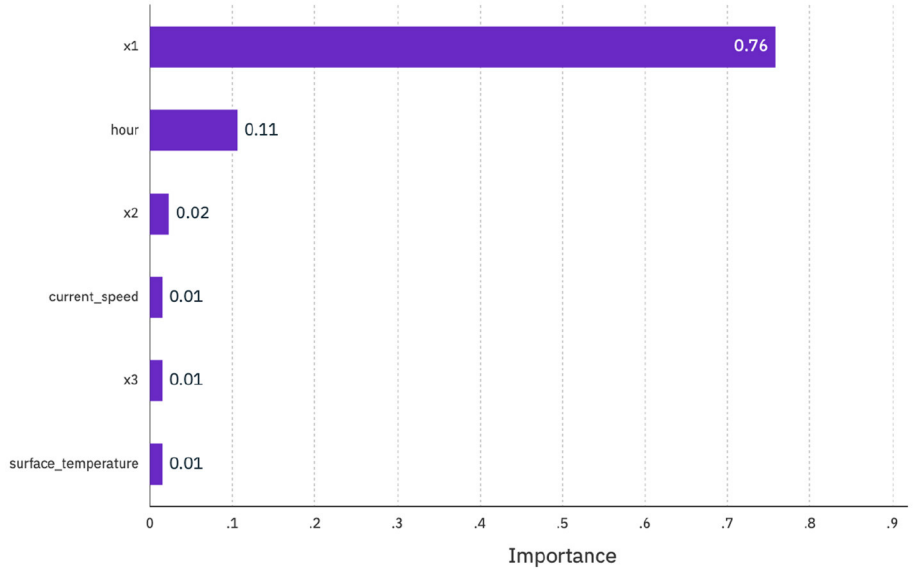

FIGURE 7 | Feature importance reported for first experimental configuration "all features" (section 2.3) at the CAN site. We provided all available data as features to the model (i.e., environmental data, temporal data, and autoregressive data or fish position at previous 3 h). The *y*-axis reports ranked list of features that contributed the most to variation in fish depth measurements, while the *x*-axis presents relative magnitude of those contributions. Ranking predictors in this manner can quickly help sift through large datasets and understand data trends (Kuhn and Johnson, 2013).

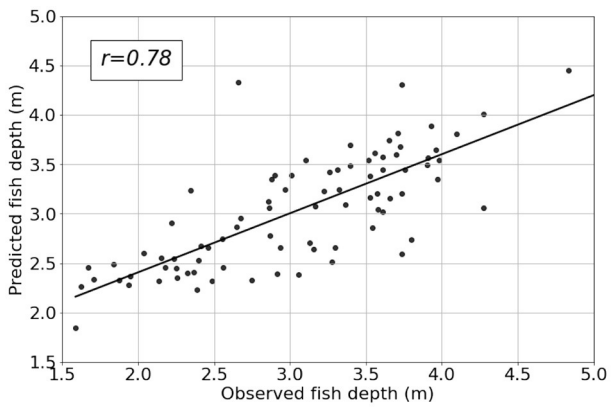

FIGURE 8 | Scatter plot of model predicted fish depth plotted against observed values for the CAN site. Inputs to the model are environmental data time-aligned with the target data, and hour of day to represent temporal variations (which **Figure 7** suggested to be an important contributor to fish motion).

sister industry, agriculture, is in the need for sensing of the ambient environment also (e.g., water temperature, oxygen)—a consideration that is less important in agriculture where animals can be housed (O'Donncha and Grant, 2019). In this paper, we adopted acoustic measurements of fish distribution to quantify how environmental conditions influence and modify behaviour.

Results demonstrated pronounced temporal variations in fish distribution as dictated by factors such as diurnal patterns, dynamics (currents and winds), and oxygen and temperature variations. Diurnal patterns driven by natural changes in light intensity were broadly similar across sites (although lack of nighttime data at SCO site limited interpretation for this site). Generally, fish occupied a deeper position in the cage during the day and were more tightly clustered; while at night, fish utilised more of the cage volume and were at a higher average position. These patterns were more pronounced at the CAN site, while the effect of longer daylight hours possibly ameliorate this effect during the June and July months at the NOR site. These diurnal patterns reflect what has been observed in the literature for salmon group response to natural light (Oppedal et al., 2011).

Analysis indicates that temperature was a primary contributor at the NOR and SCO site, while less influential at the CAN site (**Figure 9**). These results are partly influenced by the longer study period in these two sites that captured seasonal variation of temperature. **Figure 5** shows that temperature variation at the CAN site was between approximately 12–14°C compared to 10–16°C at the other two sites. Further, temperature in the warmer summer months exhibited pronounced stratification before returning to a well-mixed temperature profile in September and October. **Figure 11** presents vertical temperature profile for the NOR site, illustrating this summer stratification. Literature suggests that salmon prefer the highest available temperature (≤18°C) and avoid colder temperatures (Oppedal et al., 2001; Johansson et al., 2009). On the other hand, in reasonably homogeneous environments where temperature varies little with depth (such as CAN site during autumn), temperature is not expected to influence the vertical distribution of salmon (Oppedal et al., 2011). Hence for the sites studied here, one may expect active behavioural thermoregulation during the summer and not in other months where temperature varies little within the cage.

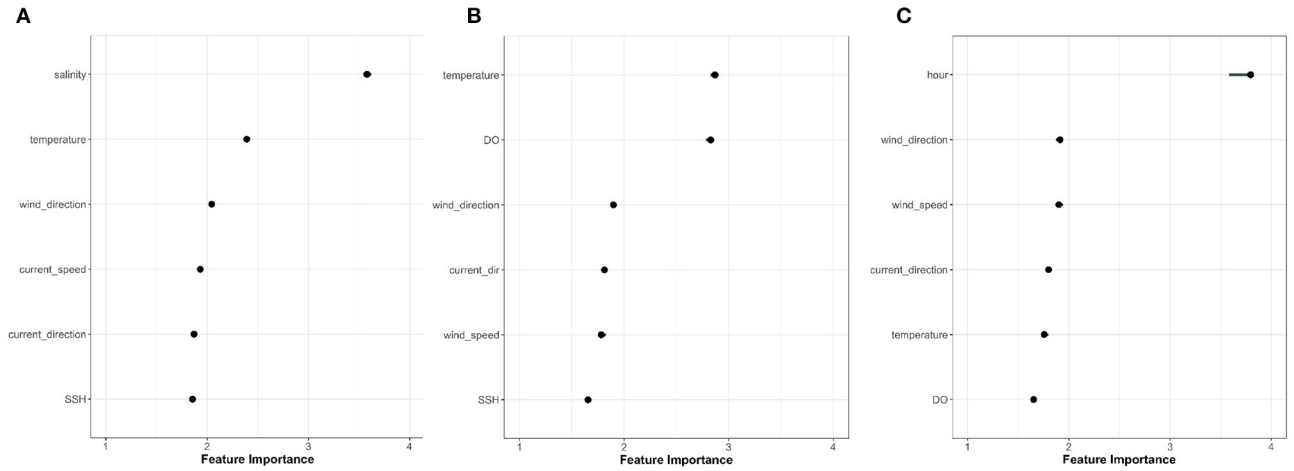

FIGURE 9 | Feature importance reported for our second experimental configuration "selected features" (section 2.3) at the **(A)** NOR, **(B)** SCO, and **(C)** CAN sites. The *y*-axis reports the ranked list of features that contributed the most to variation in fish depth measurements, while the *x*-axis presents relative magnitude of those contributions. Ranking predictors in this manner can quickly help sift through large datasets and understand data trends (Kuhn and Johnson, 2013).

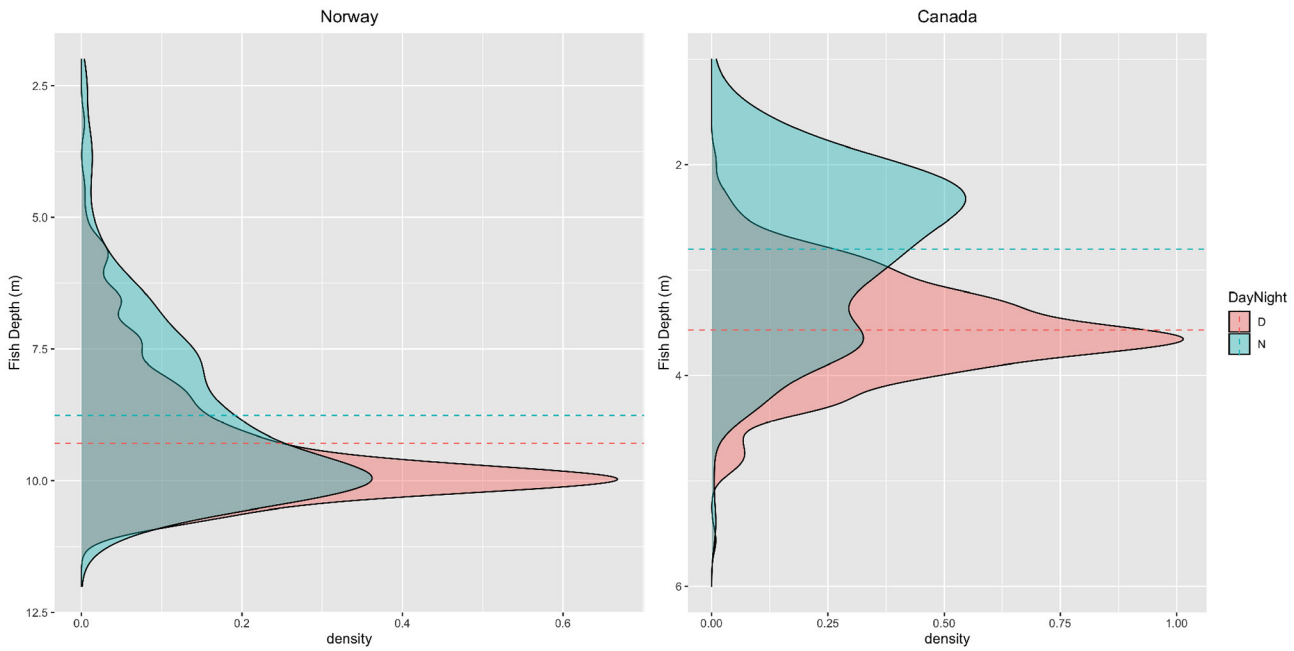

FIGURE 10 | Distribution of fish depth data for the NOR **(left)** and CAN **(right)** farm over the duration of the study period. The data is split into daytime and nighttime periods to explore how behaviours vary between those periods. The dashed vertical lines denote the mean for both periods.

Variation in oxygen levels were most pronounced at the CAN site which showed consistently lower values than at other locations (**Figure 5**). This is reflected in the feature importance analysis which denotes DO as an important contributor to fish position, and in particular indicates that lower values of oxygen have significant influence. **Figure 12** suggests that fish moved toward the surface when values drop below 7 mgL^{-1}. Values dropped below this threshold three times during the course of this study (**Figure 5**). This suggests that these low oxygen periods are worthy of additional study to understand how fish welfare were impacted and if additional behavioural modifications (e.g., horizontal swimming patterns or feeding activity) developed during these times. Research studies indicate that (at temperature of 16° C oxygen levels of 7mgL^{-1} lead to reduced appetites in full-feeding Atlantic salmon, while values of 6mgL^{-1} initiated acute anaerobic metabolism, and increased skin lesions (CREATE, 2008; Oppedal et al., 2011).

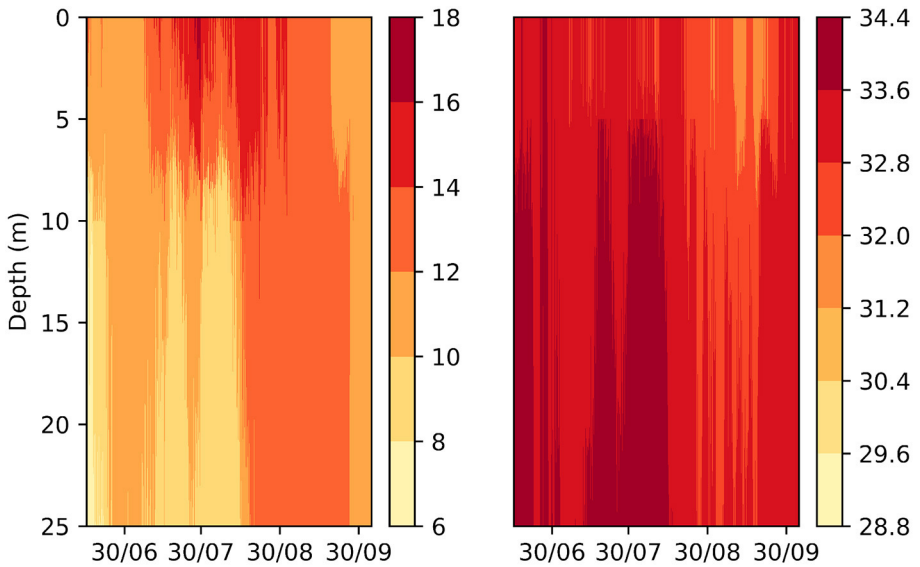

FIGURE 11 | Transect of temperature **(left)** and salinity **(right)** extracted from the Norkyst ocean model at the grid cell closest to the NOR farm. Colorbar denotes temperature in $^{\circ}C$ and salinity in units of PPU for the respective plots.

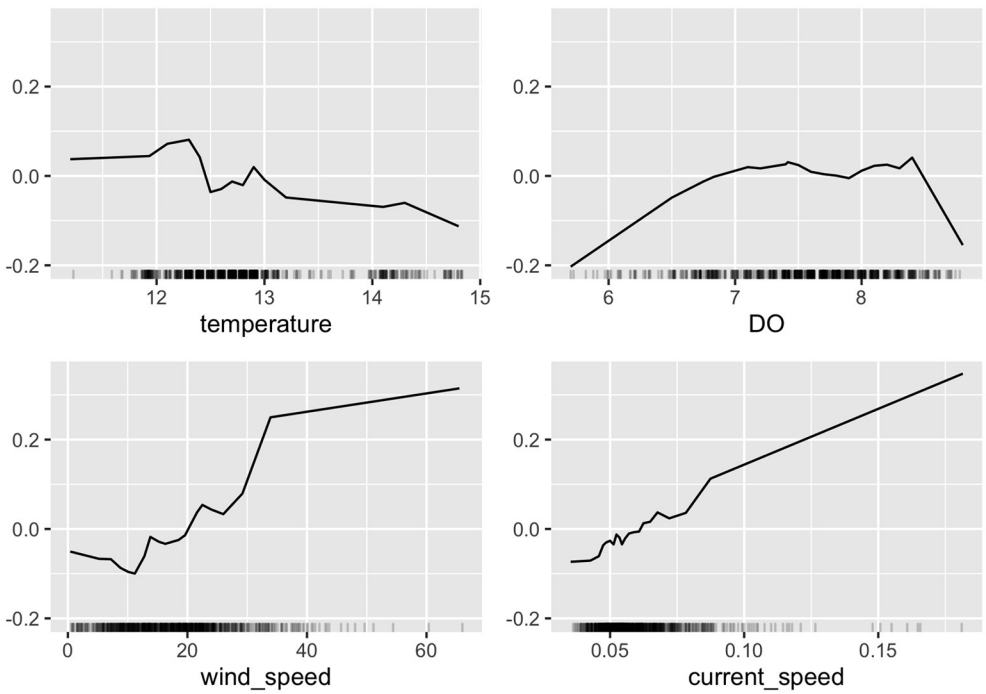

FIGURE 12 | Feature importance reported by the AutoAI model at the CAN site for a subset of environmental variables using *accumulated local effects* computation. ALE provides an efficient way to explore how much the target variable (fish depth) in response to selected feature (temperature, DO, wind speed, and current speed). The y-axis report the accumulated local effects (ALE) of each feature or variable in the units of the prediction variable (*m*), while the x-axis reports the range of values for each selected feature. Clockwise from top left results are presented for temperature, DO, current speed, and wind speed. ALE is a valuable technique to explore *how much* predictand varies in response to changes in predictors.

As alluded to in previous paragraphs, our approach only considered group behavioural responses in the vertical. Salmon typically form a circular swimming patterns that avoids both the innermost part and edges of the cage. These patterns breakdown at low stocking density, during feeding, at nighttime, or when threatened by a predator (Oppedal et al., 2011). Further, there

are important interactions that happen at the individual level, such as aggression, that are not captured here. Aggression has been shown to vary as a function of stocking density and during feeding times (Adams et al., 2007), but it is not possible to resolve at the group level. While there is potential to leverage computer vision technology to monitor at the three-dimensional (Deakin et al., 2019), or individual level (Tidal, 2020), these are currently at a laboratory or research stage. An alternative approach is to tag individual fish to collect continuous data on the three-dimensional positioning of individual fish within a cage (Roy et al., 2014). These provide very high temporal resolution of position but only a small subset of the total fish in the cage can be feasibly tracked.

Approaches based on sampling at the individual level bear similarities to more established farm management practises in agriculture. Farmers have begun to use RFID systems to track livestock movement and health in order to improve health and welfare of terrestrial livestock. These systems have provided health information such as internal body temperature, growth performance, and even hold medical information; they have also provided movement data that provides information on behaviour and interaction between individuals (Ruiz-Garcia and Lunadei, 2011). Aquaculture faces similar challenges as livestock agriculture, and therefore lessons can be adapted and applied. However, additional challenges arise when animals can move in three-dimensions, and environmental conditions affect health and welfare more consistently than in agriculture. Fish are much more dependent on farmers for food, population density, and environmental conditions (Føre et al., 2018). Further, agriculture make wide use of radio frequency communication methodologies that are not feasible underwater and instead must rely on acoustic communication channels that are less technologically mature (Stojanovic and Preisig, 2009). One of the most striking differences between both industries is that livestock farming has been occurring for millennia, where salmon farming has only been active for the last few decades, and therefore new methods and technologies to understand animal health and welfare becomes a more challenging task and requires diligent research and cooperation between farmers and researchers.

In this paper, we explored statistical and machine learning approaches to explore environmental drivers of fish behaviour using environmental observations and hydroacoustic sensors. A major challenge in research such as this is the data collection step and this varies depending on the location. In CAN, farm sites tend to be relatively shallow, ranging between 10–15 m, and in remote coastal communities. Hydroacoustic systems are successful in collecting data on high density fish populations when cage depths allow for full view. In CAN, CageEye was unable to provide a full view of the cage due to the shallowness with only 11 m of depth available when the system was designed to be placed nearer to 15–20m (pers. comm). Without the ability to place a hydroacoustic system deeper in the water column, only a small view is available. Data quality issues can be exacerbated as a result of acoustic interference from other instruments within the cage and reflection from sea surface or site bottom. Furthermore, in order to study fish behaviour, a 24-h view is needed requiring consistent clean power to be run throughout the site. With farms in more remote locations, many sites use gas-powered generators which often shut down and require maintenance as well as constant observation. In order to successfully run hydroacoustic systems at shallow sites, a clean consistent power source that can provide energy for 24-h a day would allow for uninterrupted data collection.

Results presented in this paper indicate pronounced differences between sites and the need to consider these variations for farm management. One could readily use this approach to quantify the difference between sites, and further to identify the fundamental drivers to these variations. This could be particularly valuable when comparing different farm systems such as inshore and offshore and the associated operational implications.

The primary advantage of the hydroacoustic datasets presented here are the relative ease of collection of high-density measurements of fish behaviour. This paper presents a framework to identify the dominant environmental variables influencing fish behaviour (i.e., vertical motion), and extract insight on how changes in the environment affect fish response (e.g., **Figure 12**). On the other hand, these datasets only serve as a proxy for key performance metrics that might be collected on farms, such as feeding activity or satiation, fish health as measured by things such as gill status or lice count, and mortalities. In follow-on work we will explore whether welfare indices collected on farms can be explained or predicted by a combination of sensor datasets (hydroacoustic measurements and environmental observations). In particular, we will investigate how relatively high-density, population level measurements such as hydroacoustic data can inform more sparse individual-level measurements such as sea-lice, gill health, mortalities, etc.). Further, this study considered fish behaviour in terms of group vertical movement patterns. In subsequent work, we will deploy fish tags to monitor individual fish in three-dimensions to better encapsulate individual movement patterns in three dimensions.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

FO'D, SP, and JG contributed to conception and design of the study. CS, GM, and CW collected and organised the data. FO'D wrote the first draft of the manuscript. CS, SP, GM, and RF wrote sections of the manuscript. FO'D and PP performed the statistical analysis. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This project has received funding from the European Unions Horizon 2020 research and innovation programme as part of the RIA GAIN project under grant agreement No. 773330.

REFERENCES

ABM (2020). *Aquaculture Biomass Monitor*. Available online at: https://www.biosonicsinc.com/products/aquaculture-biomass-monitor/ (accessed July 13, 2021).

Adams, C., Turnbull, J., Bell, A., Bron, J., and Huntingford, F. (2007). Multiple determinants of welfare in farmed fish: stocking density, disturbance, and aggression in *Atlantic salmon* (*Salmo salar*). *Can. J. Fish. Aquat. Sci.* 64, 336–344. doi: 10.1139/f07-018

Albretsen, J., Sandvik, A. D., and Asplin, L. (2011). *Norkyst-800: A high-Resolution Coastal Ocean Circulation Model for Norway*. Technical report, Institute of Marine Research.

Alonso, J., Villa, A., and Bahamonde, A. (2015). Improved estimation of bovine weight trajectories using support vector machine classification. *Comput. Electr. Agric.* 110:36–41. doi: 10.1016/j.compag.2014.10.001

Apley, D. W., and Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *J. R. Stat. Soc. Ser. B* 82, 1059–1086. doi: 10.1111/rssb.12377

Atoum, Y., Srivastava, S., and Liu, X. (2014). Automatic feeding control for dense aquaculture fish tanks. *IEEE Signal Proc. Lett.* 22, 1089–1093. doi: 10.1109/LSP.2014.2385794

Bjordal, Å., Juell, J., Lindem, T., and Fernö, A. (1993). "Hydroacoustic monitoring and feeding control in cage rearing of *Atlantic salmon* (*Salmo salar L.*)," in *Fish Farming Technology* (Oxfordshire), 203–208.

Boswell, K. M., Wilson, M. P., and Wilson, C. A. (2007). Hydroacoustics as a tool for assessing fish biomass and size distribution associated with discrete shallow water estuarine habitats in Louisiana. *Estuaries Coasts* 30, 607–617. doi: 10.1007/BF02841958

Boulesteix, A.-L., Janitza, S., Kruppa, J., and König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscip. Rev. Data Min. Knowl. Dis.* 2, 493–507. doi: 10.1002/widm.1072

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Bui, S., Oppedal, F., Stien, L., and Dempster, T. (2016). Sea lice infestation level alters salmon swimming depth in sea-cages. *Aquaculture Environ. Interact.* 8, 429–435. doi: 10.3354/aei00188

Buschmann, A. H., Riquelme, V. A., Hernández-González, M. C., Varela, D., Jiménez, J. E., Henríquez, L. A., et al. (2006). A review of the impacts of salmonid farming on marine coastal ecosystems in the southeast Pacific. *ICES J. Mar. Sci.* 63, 1338–1345. doi: 10.1016/j.icesjms.2006.04.021

CageEye (2021). *CageEye*. Available online at: https://www.cageeye.no/en/ (accessed July 13, 2021).

Castanheira, M. F., Conceção, L. E., Millot, S., Rey, S., Bégout, M.-L., Damsgård, B., et al. (2017). Coping styles in farmed fish: consequences for aquaculture. *Rev. Aquaculture* 9, 23–41. doi: 10.1111/raq.12100

Costa-Pierce, B. A. (2003). The 'Blue Revolution'-Aquaculture Must Go Green. *World Aquaculture* 33:4–5. doi: 10.1016/S0044-8486(02)00537-9

CREATE (2008). *Creating Aquaculture for the Future*. CREATE Annual Report 2008. Technical Report, SINTEF.

Damsgård, B., Evensen, T. H., Øverli, Ø., Gorissen, M., Ebbesson, L. O., Rey, S., et al. (2019). Proactive avoidance behaviour and pace-of-life syndrome in *Atlantic salmon*. *R. Soc. Open Sci.* 6, 181859. doi: 10.1098/rsos.181859

Deakin, A. G., Spencer, J. W., Cossins, A. R., Young, I. S., and Sneddon, L. U. (2019). Welfare challenges influence the complexity of movement: fractal analysis of behaviour in zebrafish. *Fishes* 4, 8. doi: 10.3390/fishes4010008

Dempster, T., Wright, D., and Oppedal, F. (2016). *Identifying the Nature, Extent and Duration of Critical Producation Periods for Atlantic salmon in Macquarie Harbour, Tasmania, During Summer*. Fisheries Research and Development Corporation Report, 1–16.

Drori, I., Krishnamurthy, Y., Rampin, R., Lourenço, R., One, J., Cho, K., et al. (2018). "AlphaD3M: Machine learning pipeline synthesis," in *AutoML Workshop at ICML* (Stockholm), 1–8.

FAO (2020). *The State of World Fisheries and Aquaculture 2020*. Rome: Sustainability in Action.

Fernö, A., Huse, I., Juell, J.-E., and Bjordal, Å. (1995). Vertical distribution of *Atlantic salmon* (*Salmo solar L.*) in net pens: trade-off between surface light avoidance and food attraction. *Aquaculture* 132, 285–296. doi: 10.1016/0044-8486(94)00384-Z

Ferreira, J. G., Grant, J., Verner-Jeffreys, D. W., and Taylor, N. G. H. (2013). "Carrying capacity for aquaculture, modeling frameworks for determination of," in *Sustainable Food Production* (New York, NY: Springer), 417–448.

Foote, K. G. (2009). "Acoustic methods: brief review and prospects for advancing fisheries research," in *The Future of Fisheries Science in North America* (Heidelberg: Springer), 313–343.

Føre, M., Frank, K., Dempster, T., Alfredsen, J., and Høy, E. (2017). Biomonitoring using tagged sentinel fish and acoustic telemetry in commercial salmon aquaculture: a feasibility study. *Aquaculture Eng.* 78:163–172. doi: 10.1016/j.aquaeng.2017.07.004

Føre, M., Frank, K., Norton, T., Svendsen, E., Alfredsen, J. A., Dempster, T., et al. (2018). Precision fish farming: a new framework to improve production in aquaculture. *Biosyst. Eng.* 173, 176–193. doi: 10.1016/j.biosystemseng.2017.10.014

Gebbers, R., and Adamchuk, V. I. (2010). Precision agriculture and food security. *Science* 327, 828–831. doi: 10.1126/science.1183899

Gokaraju, B., Durbha, S. S., King, R. L., and Younan, N. H. (2011). A machine learning based spatio-temporal data mining approach for detection of harmful algal blooms in the Gulf of Mexico. *IEEE J. Select. Top. Appl. Earth Observ. Remote Sens.* 4, 710–720. doi: 10.1109/JSTARS.2010.2103927

Huntingford, F., Jobling, M., and Kadri, S. (2011). *Aquaculture and Behavior*. Oxford: Wiley.

IBM (2021a). *IBM Watson Studio-AutoAI*. Available online at: https://www.ibm.com/ie-en/cloud/watson-studio/autoai (accessed July 13, 2021).

IBM (2021b). *Weather Company Data Packages*. Available online at: https://www.ibm.com/products/weather-company-data-packages (Accessed July 13, 2021).

Innovasea (2021). *Environmental Monitoring*. Available online at: https://www.innovasea.com/aquaculture-intelligence/environmental-monitoring/wireless-sensors/ (accessed July 13, 2021).

Johannesen, Á., Patursson, Ø., Kristmundsson, J., Dam, S. P., and Klebert, P. (2020). How caged salmon respond to waves depends on time of day and currents. *PeerJ* 8:e9313. doi: 10.7717/peerj.9313

Johansson, D., Laursen, F., Fernö, A., Fosseidengen, J. E., Klebert, P., Stien, L. H., et al. (2014). The interaction between water currents and slamon swimming behavior in sea cages. *PLoS ONE* 9:e97635. doi: 10.1371/journal.pone.0097635

Johansson, D., Ruohonen, K., Juell, J.-E., and Oppedal, F. (2009). Swimming depth and thermal history of individual *Atlantic salmon* (*Salmo salar L.*) in production cages under different ambient temperature conditions. *Aquaculture* 290, 296–303. doi: 10.1016/j.aquaculture.2009.02.022

Johansson, D., Ruohonen, K., Kiessling, A., Oppedal, F., Stiansen, J.-E., Kelly, M., et al. (2006). Effect of environmental factors on swimming depth preferences of *Atlantic salmon* (*Salmo salar L*) and temporal and spatial variations in oxygen levels in sea cages at a fjord site. *Aquaculture* 254, 594–605. doi: 10.1016/j.aquaculture.2005.10.029

Juell, J., Furevik, D., and Bjordal, Å. (1993). Demand feeding in salmon farming by hydroacoustic food detection. *Aquacult. Eng.* 12, 155–167. doi: 10.1016/0144-8609(93)90008-Y

Juell, J.-E., and Fosseidengen, J. E. (2004). Use of artificial light to control swimming depth and fish density of *Atlantic salmon* (*Salmo salar*) in production cages. *Aquaculture* 233:269–282. doi: 10.1016/j.aquaculture.2003.10.026

Juell, J.-E., and Westerberg, H. (1993). An ultrasonic telemetric system for automatic positioning of individual fish used to track *Atlantic salmon* (*Salmo salar L.*) in a sea cage. *Aquacult. Eng.* 12, 1–18. doi: 10.1016/0144-8609(93)90023-5

Kato, S., Nakagawa, T., Ohkawa, M., Muramoto, K., Oyama, O., Watanabe, A., et al. (2004). A computer image processing system for quantification of zebrafish behavior. *J. Neurosci. Methods* 134, 1–7. doi: 10.1016/j.jneumeth.2003.09.028

Kuhn, M., and Johnson, K. (2013). *Applied Predictive Modeling, Vol. 26*. New York, NY: Springer.

Laird, L. M. (1996). "History and applications of salmonid culture," in *Principles of Salmonid Culture, Vol. 29* (Amsterdam: Elsevier), 1–28.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., and Bochtis, D. (2018). Machine learning in agriculture: a review. *Sensors* 18:2674. doi: 10.3390/s18082674

Lindem, T., and Houari, D. A. (1993). "Hydroacoustic monitoring of fish in aquaculture-a method for automatic feeding control by detection of fish behavior," in *ICES Statutory Meeting* (Dublin).

Marine Scotland Science (2018). *Marine Scotland Science: Scottish Fish Farm Production Survey*. Available online at: https://www.gov.scot/publications/scottish-fish-farm-production-survey-2018/ (accessed July 13, 2021).

Martins, C. I., Galhardo, L., Noble, C., Damsgård, B., Spedicato, M. T., Zupa, W., et al. (2011). Beahvioural indicators of welfare in farmed fish. *Fish Physiol. Biochem*. 38, 17–41. doi: 10.1007/s10695-011-9518-8

Matthews, S. G., Miller, A. L., PlÖtz, T., and Kyriazakis, I. (2017). Automated tracking to measure behavioural changes in pigs for health and welfare monitoring. *Sci. Rep*. 7, 1–12. doi: 10.1038/s41598-017-17451-6

Norwegian Meteorological Institute (2021). *NorKyst800 Model*. Available online at: https://thredds.met.no/thredds/catalog.html (accessed July 13, 2021).

O'Donncha, F., and Grant, J. (2019). Precision aquaculture. *IEEE Intern. Things Mag*. 2, 26–30. doi: 10.1109/IOTM.0001.1900033

O'Donncha, F., and Palmes, P. (2021). *Precision Aquaculuture*. Available online at: https://github.com/IBM/PrecisionAquaculture.j.

Oldham, T., Oppedal, F., and Dempster, T. (2018). Cage size affects dissolved oxygen distribution in salmon aquaculture. *Aquacult. Environ. Interact*. 10:149–156. doi: 10.3354/aei00263

Oppedal, F., Dempster, T., and Stien, L. H. (2011). Environmental drivers of *Atlantic salmon* behaviour in sea-cages: a review. *Aquaculture* 311, 1–18. doi: 10.1016/j.aquaculture.2010.11.020

Oppedal, F., Juell, J.-E., Tarranger, G., and Hansen, T. (2001). Artificial light and season affects vertical distribution and swimming behaviour of post-smolt *Atlantic salmon* in sea cages. *J. Fish Biol*. 58, 1570–1584. doi: 10.1111/j.1095-8649.2001.tb02313.x

OxyGuard International (2014). *OxyGuard Handy Polaris 2 Portable DO Meter*. Available online at: https://www.oxyguard.dk/wp-content/uploads/2014/08/H01P-handy-Polaris-2-brochure-GB-2014-04.pdf (accessed July 13, 2021).

Palmes, P. (2020). AutoMLPipeline: A Toolbox for Building ML Pipelines. doi: 10.5281/zenodo.3980593

Palmes, P., Ploennigs, J., and Brady, N. (2020). "TSML (Time Series Machine Learning). *Proc. JuliaCon Conf*. 1, 51. doi: 10.21105/jcon.00051

Pérez-Escudero, A., Vicente-Page, J., Hinz, R. C., Arganda, S., and De Polavieja, G. G. (2014). idtracker: tracking individuals in a group by automatic identification of unmarked animals. *Nat. Methods* 11, 743–748. doi: 10.1038/nmeth.2994

Planet Tracker (2021). *Seafood Tracker Initiative*. Available online at: https://planet-tracker.org/tracker-programmes/oceans/seafood/ (accessed July 13, 2021).

Probst, P., Wright, M. N., and Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdiscip. Rev. Data Min. Knowl. Dis*. 9:e1301. doi: 10.1002/widm.1301

Roy, R., Beguin, J., Argillier, C., Tissot, L., Smith, F., Smedbol, S., et al. (2014). Testing the VEMCO Positioning System: spatial distribution of the probability of location and the positioning error in a reservoir. *Anim. Biotelem*. 2, 1–7. doi: 10.1186/2050-3385-2-1

Ruiz-Garcia, L., and Lunadei, L. (2011). The role of RFID in agriculture: Applications, limitations and challenges. *Comput. Electr. Agric*. 79, 42–50. doi: 10.1016/j.compag.2011.08.010

Saberioon, M., Gholizadeh, A., Cisar, P., Pautsina, A., and Urban, J. (2017). Application of machine vision systems in aquaculture with emphasis on fish: state-of-the-art and key issues. *Rev. Aquacult*. 9, 369–387. doi: 10.1111/raq.12143

Scherelis, C., Penesis, I., Hemer, M. A., Cossu, R., Wright, J. T., and Guihen, D. (2020). Investigating biophysical linkages at tidal energy candidate sites; A case study for combining environmental assessment and resource characterisation. *Renew. Energy* 159, 399–413. doi: 10.1016/j.renene.2020.05.109

Simmonds, J., and MacLennan, D. N. (2008). *Fisheries Acoustics: Theory and Practice*. Hoboken, NJ: John Wiley & Sons.

Stojanovic, M., and Preisig, J. (2009). Underwater acoustic communication channels: Propagation models and statistical characterization. *IEEE Commun. Mag*. 47, 84–89. doi: 10.1109/MCOM.2009.4752682

Tidal (2020). *Introducing Tidal*. Available online at: https://blog.x.company/introducing-tidal-1914257962c3 (accessed July 13, 2021).

Tonani, M., Sykes, P., King, R. R., McConnell, N., Péquignet, A.-C., O'Dea, E., et al. (2019). The impact of a new high-resolution ocean model on the Met Office North-West European Shelf forecasting system. *Ocean Sci*. 15:1133–1158. doi: 10.5194/os-15-1133-2019

Wang, D., Ram, P., Weidele, D. K. I., Liu, S., Muller, M., Weisz, J. D., et al. (2020). "AutoAI: automating the end-to-end AI lifecycle with humans-in-the-loop," in *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion* (Cagliari), 77–78.

Wolff, S., O'Donncha, F., and Chen, B. (2020). Statistical and machine learning ensemble modelling to forecast sea surface temperature. *J. Mar. Syst*. 208:103347. doi: 10.1016/j.jmarsys.2020.103347

Wolpert, D. H., and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Trans. Evolut. Comput*. 1, 67–82. doi: 10.1109/4235.585893

Conflict of Interest: FO'D and PP were employed by the company IBM Research and CW was employed by the company Cooke Aquaculture.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.