Discussion

Overall, the initial goal of the project was not achieved. Although the fit models described observed years well, they could not forecast unobserved years. Progress was made in the definition of methods to be used for the application of Machine Learning models to ecological datasets and understanding the implications and follow-ups required to fulfil the goal. This is not unusual in the science of forecasting as highly accurate forecasts that we now take for granted, such as weather forecasting, had long histories of trial and error preceding the now-reliable forecasts that are easily and widely accessible (Serafin and Wilson 2000). In order to ensure that future trials of forecasting sea lice will build on these results, specific areas that require more research should be identified. Before attempting further forecasting on this particular system, some characteristics of the system and data should be explored such as the scale of the processes, the relevance of the collected variables in forecasting, and the system's inherent potential to be forecasted.

A concern throughout the modelling of this system was that of the frequency scale of the forecast and data. Input frequencies in the data sources ranged from daily to monthly, thus forecasting on a weekly basis was done to allow for this variance. It was understood that the down-sampling and imputation of values could have had unintended consequences on the trends in the data, but the main concern regarding timescale is of the forecasted processes. The natural history of sea lice is well described in terms of generation time, with generation times ranging from eight to sixteen weeks depending on temperature (Peacock et al. 2019). Because of this, the weekly frequency of forecasts was assumed to be suitable to the system. However, within this generation time the sea lice undergo several life stages. The major distinctions between life stages that were used for this study were between motile and non-motile sea lice, which is a

distinction that is made in much of the sea lice literature, but perhaps distinction between more life stages would have uncovered interactions that were not able to be described given the current variables.

Another concern in scale comes from the physical scale at which the system was modelled. When making forecasts, only the farms around the Broughton Archipelago were included as covariates in order to reduce the number of predictors used to fit the data. These farms acted as a standing stock of disease pressure, as past studies have shown a high correlation between local farms' sea lice abundance and wild sea lice abundance (Krkošek et al. 2011). Future modelling efforts should also explore the possibility of a larger regional scale effect. Sea lice can disperse on very large scales, and so models predicting the abundance of sea lice should also explore explicitly modelling the dispersal of lice on larger scales than that explored in this study.

Machine learning methods were chosen for this particular system because relative to many other wildlife disease datasets, the dataset available for this system is quite large and comprehensive. Bringing together all the data sources, the data used in this project described the abundance of sea lice in wild salmon populations for over 15 years and the abundance of sea lice in farmed salmon for over 7 years. Typically, machine learning performs best on large amounts of data and so this system seemed like a good fit for the methods used. Given the relative amount of data available here, one might think that machine learning is the obvious choice for describing this system, but this may be too narrow a point of view.

Many forecasting efforts in ecology do not rely on machine learning for their predictions, but on mathematical process models (Houlahan et al. 2017, Dietze et al. 2018). Machine learning, when found in the disease forecasting models, is mostly used to estimate parameters

and ensemble predictions from mathematical models rather than to produce predictions on its own. In the context of disease, it is interesting to make the comparison in the suitability of machine learning algorithms in modelling systems versus mathematical process models.

It is easy to see how a mathematical model is well-suited for describing a disease such as SARs (or more relevant in 2020, COVID-19); it is conceptually simple to break up the population in an area into susceptible, exposed, infected, etc. and from there to model these different groups. External variables might be less important than the contact between individuals and size of groups when making predictions on the disease's outcome and these kinds of transmission models have been extensively studied. However, in a system such as sea lice, the suitability of a mathematical model is less obvious. Macroparasites such as sea lice are much less studied and modelled and so a robust process model such as SEIR is not available for sea lice. The modelling done in this project was done with the assumption that external variables such as temperature and salinity would have a major impact on the abundance of sea lice in a given year, but the results from this study did not find these factors to have predictive power. Perhaps given the same data, a mathematical model could have been parametrised to produce accurate forecasts, however it is not possible to answer this question without speculation as it would require going back to the beginning of the project and exploring a different route of modelling. The suitability of mathematical versus machine learning models for this particular system is still undetermined, and further research should explore the advantages and disadvantages that each method provides.

Overall, more research should be done into the forecasting of sea lice. Starting with this project as a base it would be reasonable to add more data from different sources, explore the

scale of the system further, and compare different modelling approaches in order to determine if useful insights can be achieved.