It's raining sea lice: forecasting sea lice occurrence in juvenile wild salmon

Tristan Garry

Supervised by Marie-Josée Fortin and Martin Krkošek

Introduction

In a world increasingly marked by big data and easy access computing, scientific disciplines have been able to take advantage of analytical advances from multiple disciplines. Ecology is becoming a more predictive science (Dietze et al. 2018, White et al. 2019) and is moving towards using more computationally intensive methods (Green et al. 2005). With widespread access to large datasets and stronger computing, forecasting is becoming easier to perform and as such our perspective on the subject is changing. Ecology has in the past lacked short-term iterative forecasts, a form of forecasting performed repeatedly on relatively short timescales with consistent evaluation (Dietze et al. 2018, White et al. 2019). Without iterative forecasting on new data, it is difficult for us to evaluate our ability to make causal inferences and our ability to improve our understanding of our forecasting ability is hindered. There is little literature on forecasting in ecology compared to other fields, perhaps due to the perceived difficulty of accurately modelling biological systems (Houlahan et al. 2017).

Forecasting is imperative in many fields and industries such as economics (Teschner and Weinhardt 2015) and energy (Suganthi and Samuel 2012). Such fields have seen the use of vast amounts of different modelling techniques in order to best accommodate the high demand for forecasts; ecology has not seen the same magnitude of forecasting efforts. Themes applicable to ecology have been explored in forecasting population behaviour in macroeconomics (Teschner and Weinhardt 2015), to forecasting disease in public health (Shaman and Karspeck 2012). It has

been found in numerous instances that repetition of forecasts over time improves their accuracy (Teschner and Weinhardt 2015, Dietze et al. 2018), and fields such as economics and energy have benefited greatly from large amounts of repetition and review. While ecology does not have the same history in performing forecasting as many fields and industries, it would benefit the field of ecology to look to other fields when further developing its own forecasting expertise.

Forecasting has been applied in ecology for forecasting populations in fisheries management, the conservation of threatened species, and examining population responses to perturbations (Ward et al. 2014). The intricacies of ecological data, such as population and spatial structure, has brought on the adoption of some novel modelling techniques in the field. As an example, ecologists have had to develop models able to independently define subpopulations by using methods such as state-space models (Ward et al. 2010) in order to account for spatial structure. Although this complexity may seem daunting at certain levels of ecological organisation (Ward et al. 2014), many attempts at forecasting populations have been successful and iterating on these forecasting approaches is key in order to move our understanding of the predictability of ecology forward (White et al. 2019).

Autoregressive integrated moving average (ARIMA) models have been used in many applications of forecasting (Zhang 2003). ARIMA models have been used in many applications and fields such as the forecasting of fuel demands for policy making (Ediger and Akar 2007) or the prediction of mosquito population responses to changing weather (Groen et al. 2017). These types of models are very popular due to the Box-Jenkins method (Makridakis and Hibon 1997) being well-described and working well on time series. Seeing as they have been explored in many applications, these models are often considered when developing new forecasting methods and will be explored extensively in this project.

Another set of models that is of interest in forecasting is artificial neural networks (ANN) (Böse et al. 2017). This type of model "learns" relationships present in data in non-parametric ways by creating a graph of connected nodes with weights that fit to the data passed in to them in order to describe a pre-determined output. This perhaps addresses concerns that biology is "too complex" to accurately model (Dietze et al. 2018) as no functional form for the data-generating model has to be assumed. ANNs have shown very strong results in many applications such as natural language processing and image recognition (Gu et al. 2018) and have shown strong results in forecasting in certain applications (Kimoto et al. 1990). Combination of neural network and parametric methods have also seen great success (Makridakis et al. 2018) and may even be preferred in some circumstances. Advances in machine learning (Crone et al. 2011, Nevo et al. 2018, Wang et al. 2019) could be used to forecast difficult-to-model ecological relationships and expand on the growing imperative of ecological forecasting.

Forecasting has seen many uses in disease research, where an accurate forecast of an outbreak can be crucial to the health of a population (Petukhova et al. 2018). Disease can have drastic effects on population viability, as we have seen recently in White-nose syndrome in the bats of North America (Wilder et al. 2011). Of current relevance, climate change can have strong effects on the prevalence of disease (Zamora-vilchis et al. 2012) and so attention should be given to the understanding of these relationships. Parametric forecasting methods such as ARIMA models have been applied for the forecasting of many diseases historically (Kaundal et al. 2006, Molento et al. 2018). More recently, ANNs have shown to be very powerful when estimating disease response to climate due to the sometimes difficult to model relationship between disease and its environment (Kaundal et al. 2006).

A disease of particular interest to this project is the parasitism of salmon by sea lice on the west coast of Canada. Sea lice parasites have had negative effects on both wild and farmed salmon populations in the area, but parasites have shown to be controllable given the proper treatment (Peacock et al. 2013). Abundance of the lice in local fish farms has been shown to have a correlation with occurrence of the lice in wild fish (Marty et al. 2010), perhaps indicating that lice management on farms would have significant effects on wild populations. Critically, the timing of management efforts has been shown to be important in the control of sea lice abundance on salmon (Bateman et al. 2016) and so being able to forecast outbreaks would be a useful management tool. Improvement in these forecasting methods could perhaps also spur further development in the field of forecasting ecological systems.

There are two species of lice that are common on Pacific salmon: *Lepioptheirus salmonis*, a sea louse that is a salmon specialist, and *Caligus clemensi*, a generalist that is less pathogenic to salmon. Sea lice reproduce sexually and produce nauplii that are able to disperse tens of kilometres along ocean currents at the first stage of their life. These nauplii moult into copepodites, with timing depending on temperature. At this stage, the lice are infectious to fish and attach to a host before progressing through immobile chalimus stages and then two mobile pre-adult stages. At the adult stage, sea lice are mobile on the surface of their host and may move between hosts (Peacock et al. 2019). These sea lice have been the subject of extensive research and management efforts (Peacock et al. 2013). Accurate forecasting of sea lice abundance, and specifically outbreaks, could help the control of sea lice infestations in the region and help deter the damage caused by parasites on both wild and farmed animals.

Specific Aims

- Create a parametric model taking farmed and wild salmon data as well as climate data as inputs that can forecast sea lice abundance between May and July on the coast of BC
- Create a machine learning model trained using farmed and wild salmon data as well as climate data that can more accurately forecast sea lice abundance between May and July on the coast of BC than a parametric model
- Explore artificial neural network and parametric-neural network combination forecasting techniques that have not seen extensive use in ecology

Methods

Data

Wild Juvenile Salmon Data

Forecasting efforts will be focused on estimating sea lice abundance on wild juvenile salmon in the Broughton Archipelago, BC. The data to be interpreted and forecasted consist of a monitoring effort ongoing from 2001 to 2019, specific details on sampling and methods are available in a supporting paper (Peacock et al. 2016). To summarise the sampling methods, sampling consisted of visually searching coastal waters, ~2-5m from the shore, and capturing schools of juvenile salmon for examination. Once captured, the salmon were measured, and health characteristics and species were noted. The salmon were non-lethally visually examined using 16x magnification for sea lice, with life stage and species of lice being noted. As well as

fish and lice information, site information was collected for each sampling event including temperature and latitude/longitude.

Using this dataset, I will develop models to forecast the average occurrence of motile sea lice on juvenile salmon between May and July. Motile sea lice for this analysis are defined as pre-adult and adult stage male and female sea lice, and models will be trained using the motile sea lice counts calculated from this dataset. Variables of interest from this dataset include fish body size characteristics (height, length), date, site information (temperature, location), and lice counts.

Farmed Salmon Data

Parasite occurrence in farms is thought to have a significant effect on parasite occurrence on wild fish (Marty et al. 2010) and so will be analysed as a variable that could potentially influence the forecasting of parasitism in the Broughton Archipelago, BC. Two sources of data are publicly available monthly for farmed fish: Fisheries and Oceans Canada (DFO) audit data collected jointly by DFO and farm staff, and the industry counts of sea lice occurrence collected by farm staff. Both of these datasets consist of monthly averages of motile sea louse occurrence, resolution to the individual sampling event level is not publicly available. DFO audit data are less comprehensive in farm and month coverage for farms relevant to the Broughton Archipelago due to DFO audits happening less frequently than the industry counts but will likely be focused on due to concerns of under-reporting in industry counts (Godwin et al. 2019). Included in this dataset is the occurrence of treatment events for over-occurrence of sea lice, sea louse treatment is mandated by law when average counts in a farm reach above 3 per fish; these treatments have been shown to have significant effects on louse abundance (Peacock et al. 2013). Since farmed

salmon louse abundance has an effect on wild salmon louse abundance (Marty et al. 2010), we should expect these treatment events to significantly affect the Broughton Archipelago louse abundances.

Modelling

This project ultimately aims to compare two models to evaluate forecasting performance: one parametric ARIMA (Auto-regressive Integrated Moving Average)-based model, and one non-parametric ANN (Artificial Neural Network)-based model. ARIMA models have been used extensively in other forecasting related fields such as economics and meteorology, while ANN models continue to emerge as potentially very powerful alternatives in the face of rich datasets. This project also aims to expand on ANN use for forecasting in ecology while also parametrising an ARIMA model in order to compare the performance of different forecasting methods on complex ecological relationships. The first phase of this project will consist of evaluating several models' suitability in forecasting sea louse occurrence with the later stages of the project focusing on applying the models found to be appropriate for this system.

One of the model types explored will consist of an ARIMA-based model. Many variations of ARIMA models have been used in forecasting (Chen and Boccelli 2018) and so these kinds of models may be of use for describing the sea lice infections of salmon. ARIMA models have typically been used for linear relationships, and so their suitability to this particular system of study will have to be evaluated.

As ANNs have only continued to show improvements in performance, an ANN based model will be evaluated for performance on the sea lice data. Typical forecasting methods tend to be parametric and so a form has to be assumed in order to model the observed relationships.

Since ANN are non-parametric, these methods do not require an assumed form of the relationships in the data and so might be the best suited tool for modelling ecological systems given enough data. ANN forecasting methods have been shown to greatly outperform parametric methods in non-linear forecasting (Zhang and Qi 2005, Wang et al. 2019) and so will be explored extensively due to their potential performance and not having to assume linearity.

Although performing well separately on different problems, there is growing evidence that combining parametric and non-parametric methods into a single model can account for the shortcomings of both methods (Zhang 2003, Pai and Lin 2005, Khashei and Bijari 2011, Makridakis et al. 2018). As real data are often not completely linear or non-linear, the middle ground could prove to be the most effective modelling strategy.

Data Preparation

A challenge that has been encountered in the forecasting of this population is the nature of this dataset compared to the data used in the greater forecasting literature. The majority of forecasting research consists of regular time-series, that is the time step difference between data points when ordered sequentially tends to be consistent or predictable across the dataset. This is a problem for model selection as the main dataset being used for forecasting consists only of sampling done from May-July. The brokenness of this time series creates problems for conventional model application, such as the regular usage of ARIMA and some ANN models.

To tackle this problem, I will investigate two solutions. The first solution is to complement (impute) the missing time with other datasets such as the fish farm data and applying the model over the time series as conventional time series modelling is performed. This solution would require a strong relationship between the supplementary data out-of-season and

the primary data within-season and would leave the majority of forecasting methods available to use. ARIMA methods have been shown to be more sensitive to missing data than ANN methods in certain circumstances (Chen et al. 2001) so this solution could have important consequences on the performance of parametric models.

The second solution involves treating each season as a separate sample of a shorter-term, within-season forecast. Although this is the preferred application of forecasting in this circumstance the literature on within-season forecasting has been less developed, particularly in the neighbourhood of neural networks. As such, choosing this method may result in stronger performance of parametric methods while having consequences on the performance of non-parametric methods.

A benchmark performance will also be established for forecasting performance using a seasonal naïve model. This model consists simply of returning the last known observation of the same period. Seasonal naïve models typically perform well on highly seasonal data; the parasitism data can be considered seasonal when not experiencing outbreak conditions and so this performance benchmark may be reasonable as a goal to beat. Being unable to beat a seasonal naïve model would indicate that higher levels of complexity are unnecessary in the forecasting of this system and that the fit models are perhaps overcomplicating the system.

Results Interpretation

Typical forecasting metrics to indicate performance include RMSE (Residual Mean Squared Error), MAE (Mean Absolute Error), and MAPE (Mean Absolute Percentage Error). Forecasting accuracy and precision will be evaluated based on these metrics and will be compared to the benchmark naïve seasonal model.

A particular performance of interest is the models' ability to predict outbreak seasons. Parasite occurrence between farmed and wild salmon has been shown to be linked (Marty et al. 2010) and so being able to account for outbreaks in farms as well as treatment events will be very important in evaluating the performance of the forecasts. The benchmark model will perform quite poorly on the prediction of unusual seasons and so given a standard framework of accepting a model performing better than the benchmark, the bar for a well performing model may be set artificially low. Further research will be required into the evaluation of outbreak forecasting accuracy as this is an import facet of the use of these particular models.

Timeline

Month	Actions
October	- Exploration of modelling avenues applicable to the sea lice data
	- Exploration of data preparation methods
	- Proposal due (October 28th)
November	- Finalisation of data preparation methods
	- Narrow possible models to 3-4
	- Thoroughly evaluate possible models on sea lice data
December	- Select final models
	- Evaluate initial performance of models on sea lice data
January	- Fine-tuning selected models
February	- Results due (February 3rd)
	- Begin report write-up & discussion
March	- Continue write-up & discussion
	- Make poster
	- Presentation of results (March 27th)
April	- Adjust report based on presentation feedback
	- Final report due (April 3rd)

References

Bateman, A. W., S. J. Peacock, B. Connors, Z. Polk, D. Berg, M. Krkošek, and A. Morton. 2016. Recent failure to control sea louse outbreaks on salmon in the Broughton Archipelago, British Columbia. Canadian Journal of Fisheries and Aquatic Sciences 73:1164–1172.

Böse, J. H., V. Flunkert, J. Gasthaus, T. Januschowski, D. Lange, D. Salinas, S. Schelter, M. Seeger, and Y. Wang. 2017. Probabilistic demand forecasting at scale. Proceedings of the VLDB Endowment 10:1694–1705.

Chen, H., S. Grant-Muller, L. Mussone, and F. Montgomery. 2001. A study of hybrid neural network approaches and the effects of missing data on traffic forecasting. Neural Computing and Applications 10:277–286.

Chen, J., and D. L. Boccelli. 2018. Environmental Modelling & Software Real-time forecasting and visualization toolkit for multi-seasonal time series. Environmental Modelling and Software 105:244–256.

Crone, S. F., M. Hibon, and K. Nikolopoulos. 2011. Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction. International Journal of Forecasting 27:635–660.

Dietze, M. C., A. Fox, L. M. Beck-Johnson, J. L. Betancourt, M. B. Hooten, C. S. Jarnevich, T. H. Keitt, M. A. Kenney, C. M. Laney, L. G. Larsen, H. W. Loescher, C. K. Lunch, B. C. Pijanowski, J. T. Randerson, E. K. Read, A. T. Tredennick, R. Vargas, K. C. Weathers, and E. P. White. 2018. Iterative near-term ecological forecasting: Needs, opportunities, and challenges. Proceedings of the National Academy of Sciences of the United States of America 115:1424–1432.

Ediger, V. Ş., and S. Akar. 2007. ARIMA forecasting of primary energy demand by fuel in

Turkey. Energy Policy 35:1701–1708.

Godwin, S. C., M. Krkosek, J. D. Reynolds, and A. W. Bateman. 2019. Bias in self-reported parasite data from the salmon farming industry.

Green, J. L., A. Hastings, P. Arzberger, F. J. Ayala, K. L. Cottingham, K. Cuddington, F. Davis, J. A. Dunne, M.-J. Fortin, L. Gerber, and M. Neubert. 2005. Complexity in Ecology and Conservation: Mathematical, Statistical, and Computational Challenges. BioScience 55:501.

Groen, T. A., G. L'Ambert, R. Bellini, A. Chaskopoulou, D. Petric, M. Zgomba, L. Marrama, and D. J. Bicout. 2017. Ecology of West Nile virus across four European countries: Empirical modelling of the Culex pipiens abundance dynamics as a function of weather. Parasites and Vectors 10:1–11.

Gu, J., Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen. 2018. Recent advances in convolutional neural networks. Pattern Recognition 77:354–377.

Houlahan, J. E., S. T. Mckinney, T. M. Anderson, and B. J. Mcgill. 2017. The priority of prediction in ecological understanding. Oikos:1–7.

Kaundal, R., A. A. Kapoor, and G. P. S. Raghava. 2006. Machine learning techniques in disease forecasting: A case study on rice blast prediction. BMC Bioinformatics 7:1–16.

Khashei, M., and M. Bijari. 2011. A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. Applied Soft Computing Journal 11:2664–2675.

Kimoto, T., K. Asakawa, M. Yoda, and M. Takeoka. 1990. Stock market prediction system with modular neural networks. Pages 1–6 1990 IJCNN international joint conference on neural networks.

Makridakis, S., and M. Hibon. 1997. ARMA models and the Box-Jenkins methodology. Journal

of Forecasting 16:147–163.

Makridakis, S., E. Spiliotis, and V. Assimakopoulos. 2018. The M4 Competition : Results , findings , conclusion and way forward The M4 Competition : Results , findings , conclusion and way forward. International Journal of Forecasting.

Marty, G. D., S. M. Saksida, and T. J. Quinn. 2010. Relationship of farm salmon, sea lice, and wild salmon populations. Proceedings of the National Academy of Sciences of the United States of America 107:22599–22604.

Molento, M. B., S. Bennema, J. Bertot, I. C. Pritsch, and A. Arenal. 2018. Bovine fascioliasis in Brazil: Economic impact and forecasting. Veterinary Parasitology: Regional Studies and Reports 12:1–3.

Nevo, S., V. Anisimov, G. Elidan, R. El-Yaniv, P. Giencke, Y. Gigi, A. Hassidim, Z. Moshe, M. Schlesinger, G. Shalev, A. Tirumali, A. Wiesel, O. Zlydenko, and Y. Matias. 2018. ML for Flood Forecasting at Scale. Pages 2–5 32nd Conferene on Neural Information Processing Systems.

Pai, P. F., and C. S. Lin. 2005. A hybrid ARIMA and support vector machines model in stock price forecasting. Omega 33:497–505.

Peacock, S. J., A. W. Bateman, B. Connors, M. A. Lewis, and M. Krkošek. 2019. Ecology of a marine ectoparasite in farmed and wild salmon. Pages 544–573 Wildlife Disease Ecology. Cambridge University Press, Cambridge, UK.

Peacock, S. J., A. W. Bateman, M. Krkošek, B. Connors, S. Rogers, L. Portner, Z. Polk, C. Webb, and A. Morton. 2016. Sea-louse parasites on juvenile wild salmon in the Broughton Archipelago, British Columbia, Canada. Ecology 97:1887.

Peacock, S. J., M. Krkosek, S. Proboszcz, C. Orr, and M. A. Lewis. 2013. Cessation of a salmon

decline with control of parasites. Ecological Applications 23:606–620.

Petukhova, T., D. Ojkic, B. Mcewen, R. Deardon, and Z. Poljak. 2018. Assessment of autoregressive integrated moving average (ARIMA), generalized linear autoregressive moving average (GLARMA), and random forest (RF) time series regression models for predicting influenza A virus frequency in swine in Ontario , Canada. PLOS ONE:1–17.

Shaman, J., and A. Karspeck. 2012. Forecasting seasonal outbreaks of influenza. Proceedings of the National Academy of Sciences of the United States of America 109:20425–20430.

Suganthi, L., and A. A. Samuel. 2012. Energy models for demand forecasting — A review. Renewable and Sustainable Energy Reviews 16:1223–1240.

Teschner, F., and C. Weinhardt. 2015. A macroeconomic forecasting market. Journal of Business Economics 85:293–317.

Wang, Y., A. Smola, D. C. Maddix, J. Gasthaus, D. Foster, and T. Januschowski. 2019. Deep Factors for Forecasting.

Ward, E. J., H. Chirakkal, M. González-Suárez, D. Aurioles-Gamboa, E. E. Holmes, and L. Gerber. 2010. Inferring spatial structure from time-series data: Using multivariate state-space models to detect metapopulation structure of california sea lions in the gulf of California, Mexico. Journal of Applied Ecology 47:47–56.

Ward, E. J., E. E. Holmes, J. T. Thorson, and B. Collen. 2014. Complexity is costly: A meta-analysis of parametric and non-parametric methods for short-term population forecasting. Oikos 123:652–661.

White, E. P., G. M. Yenni, S. D. Taylor, E. M. Christensen, E. K. Bledsoe, J. L. Simonis, and S. K. M. Ernest. 2019. Developing an automated iterative near-term forecasting system for an ecological study. Methods in Ecology and Evolution 10:332–344.

Wilder, A. P., W. F. Frick, K. E. Langwig, and T. H. Kunz. 2011. Risk factors associated with mortality from white- nose syndrome among hibernating bat colonies. Biology Letters 7:950–953.

Zamora-vilchis, I., S. E. Williams, and C. N. Johnson. 2012. Environmental Temperature Affects Prevalence of Blood Parasites of Birds on an Elevation Gradient : Implications for Disease in a Warming Climate. PLOS ONE 7.

Zhang, G. P., and M. Qi. 2005. Neural network forecasting for seasonal and trend time series. European Journal of Operational Research 160:501–514.

Zhang, P. G. 2003. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50:159–175.