Imperial College London

IMPERIAL COLLEGE LONDON

DEPARTMENT OF LIFE SCIENCES

Deep neural networks to predict foraging behaviour: salt-water immersion data can accurately predict diving in seabirds.

Author:
Luke Swaby

CID:
01980806

Words:
5892

Date:
August 26, 2021

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science at Imperial College London
Submitted for the MSc in Computational Methods in Ecology and Evolution.

August 26, 2021

Declaration

The data reviewed in this project was collected by Robin Freeman (ZSL), who then extracted it using the XManager software (`https://www.technosmart.eu/`) and handed it over to me. I conducted all the following data processing and cleaning, as well as subsequent model building, training, and evaluation, but Robin closely advised and supervised the implementation of methods throughout.

Abstract

Identifying and protecting foraging locations is a key step in the conservation of wildlife. For elusive and wide-ranging species, this process often depends on the deployment of cumbersome, power-hungry biologging devices that can significantly alter the behaviour of the animals wearing them, placing time constraints on studies that use them and casting doubt on the validity of the data they produce. Using deep neural networks, I here evaluate the performance of salt-water immersion data derived from lightweight, 3g geolocators (GLS) as predictors of foraging behaviour in a species of pursuit-diving seabird, thereby providing the first known assessment of the potential for using GLS alone to identify foraging locations for wide-ranging species throughout their annual cycles. I use fine-scale tri-axial acceleration data as a performance benchmark and cross-validate predictions with data from withheld birds to ensure robust classification assessment. A probabilistic method of georeferencing predictions is then introduced in order to investigate the challenges of mapping these along flight paths in practise. When evaluated on reduced, balanced data sets, optimal models classified the behavioural states (dive/non-dive) of windows of data points with 98.5% and 93.67% accuracy for acceleration and immersion data respectively. Similar predictive accuracy was achieved when mapping predictions to recorded GPS coordinates. The results described in this project therefore show that GLS alone can be used to accurately identify diving events, and, by proxy, foraging locations, for pursuit-diving seabird species year-round at a substantially lower cost both to researchers and animals wearing the devices than is currently accepted with the conventional method.

1 Introduction

In the marine environment, the foraging behaviours of many top predators are dependent on the distribution and predictability of resources across large spatial scales, and are accordingly considered to encode valuable information about the health of ecosystems globally (Hazen et al., 2019, Ciancio et al., 2021, Carignan and Villard, 2002). Pelagic seabirds in particular are often proposed as indicators of ecological status, and have been used extensively to inform ecosystem management and conservation strategies (Parsons et al., 2008, Thaxter et al., 2012, Bost and Le Maho, 1993, Einoder, 2009, Mallory et al., 2010). However, with direct observation largely precluded by their elusive and wide-ranging nature, fine-scale behavioural data on these species have only become available in recent years, where innovations in telemetry and bio-logging technology have liberated researchers from the epistemic constraints of sparse presence/absence data generated by traditional surveys at sea and ringing recoveries and truly pulled back the curtain on their lives over the open ocean (Rutz and Hays, 2009, Guilford et al., 2009, Maclean et al., 2013, Bograd et al., 2010). Perhaps most notably, the advent and continually declining size and cost of global positioning systems (GPS) and time-depth recorders (TDR) has enabled the collection of a wealth of high-resolution movement data that has been especially valuable in identifying the behaviours associated with foraging for pursuit-diving seabird species (Guilford et al., 2008, Thaxter et al., 2012).

In studies investigating such behaviours to date, the method predominantly employed has been to use a combination of the two, mapping continuous dive profiles recorded by the latter to GPS coordinates recorded by the former by their respective timestamps to pinpoint dive events that can then be used as proxies to identify foraging locations (Wanless et al., 1997, Shoji et al., 2015, Dean et al., 2013). However, this technique has several limitations. One is the considerable expense of purchasing TDR devices. Another is the time-constraint imposed by the relatively short battery lives of high-sampling-rate GPS loggers, often restricting the duration of studies that use them to just a few months and additionally introducing the possibility of generating large quantities of surplus data whenever simultaneously deployed devices with differing battery lives are activated or deactivated asynchronously. As a consequence, there currently exists a substantial archive of data that remains unexamined and, *ipso facto*, untapped for potentially significant ecological insights to date.

Despite the continuing miniaturization of logging devices, it has also been reported that the increase in wing-loading caused by even lighter modern devices can significantly alter the behaviour of birds wearing them, casting doubt on the validity of the data they produce (Barron et al., 2010, Phillips et al., 2003, Calvo and Furness, 1992, Jackson and Wilson, 2002). For example, Gillies et al. (2020) found that Manx shearwaters (*Puffinus puffinus*) fitted with 17g GPS loggers (4.2% of bird body mass) more than doubled the length of their foraging trips and almost quartered the mass gained from those same trips as compared to birds carrying no device or just a small 2.5g leg-mounted device. Considering that even compact, multi-sensor loggers often still weigh over 15g, such findings strongly suggest that substantial weight reductions will still be needed before the data yielded by these devices can be interpreted as reflecting 'natural behaviour' with any confidence, especially when affixed to smaller animals or where multiple devices are used in unison.

An alternative to GPS for animal tracking is light-level geolocation (GLS). These remarkably lightweight devices (0.3g+) are equipped with wet/dry sensors as well as ambient light sensors to determine location, and are often used to map bird migration routes (Newton, 2010, Åkesson et al., 2012, Minton et al., 2010, Bächler et al., 2010, Jahn et al., 2013). The spatio-temporal resolution of the data they produce is coarse compared to GPS data ($\sim$12 hours and 186 $\pm$ 144km; Phillips et al. (2004)), but what is lost in resolution is compensated by the longevity of their battery lives, which can last for years on a single charge. As a result, vast quantities of GLS data exist for birds for whom no other data is available. This wide availability of data, in combination with the device's ultra-light weight and long operational life, makes GLS the gold standard for studies aiming to identify foraging behaviours of wide-ranging pelagics, for if it were possible to glean such information from GLS data alone, then this would allow researchers to collect and analyse substantially more data (from a broader range of species covering wider areas) at lower costs to both themselves and the animals wearing the devices than the current method permits, as well as map successful model predictions for the numerous animals for which only GLS data is available.

Traditionally, the challenges of analysing the vast and complex data sets generated by modern bio-loggers (discussed, for example, in Urbano et al. (2010)) have been tackled in behavioural studies with unsupervised approaches such as state-space models (e.g. hidden Markov Models) and Gaussian mixture models (Rutz and Hays, 2009, Dean et al., 2013, Jonsen et al., 2005, Patterson et al., 2009, Breed et al., 2012). In more recent years, some have assembled labelled data sets to leverage the abilities of more data-intensive, supervised machine learning (ML) techniques that can achieve similar predictive feats with even greater accuracy (Guilford et al., 2009, Nathan et al., 2012, Martiskainen et al., 2009, Grünewälder et al., 2012, Carroll et al., 2014). However, many of these techniques place heavy demands on the user, often requiring the manual extraction of *features* (reduced summary statistics to simplify learning) from the raw data; an often laborious procedure that requires specialised knowledge and can drastically ramp up the complexity of implementing such models.

Deep Neural Networks (DNNs) are a family of powerful, general-purpose ML models that not only excel in handling big data, but also automate the process of feature extraction, thereby enabling fine-grained analysis of raw input data without demanding any domain-specific expertise of the user (Bishop et al., 1995). By iteratively propagating data across multiple processing layers and using the resulting predictions to calculate the gradient of the network's error with respect to each of its individual parameters and adjust them accordingly (an algorithm known as 'backpropagation'), these models are capable of learning highly intricate properties of input data with many levels of abstraction. They have accordingly gained a lot of traction in the era of big data, where explosions in computational capacity and intractable data sets have highlighted their advantages over other more conventional ML methods. However, despite driving breakthroughs in (and in many cases revolutionising) a broad range of fields (see LeCun et al. (2015) for a review), they have scarcely been used in the context of animal behavioural studies to date, having often been overlooked in favour of the more traditional aforementioned behavioural prediction techniques. But while this may be sensible in cases where one wishes to understand the behavioural patterns distinctive of each class in a predictive problem (an objective precluded by the 'black-box' nature of neural networks), in the alternative case where one is concerned only with generating quick and accurate predictions, and not the mechanistic processes by which they are made, DNNs represent a powerful yet more user-friendly tool for the job.

Reinforcing this sentiment, a recent study demonstrated that DNNs can be used to predict the diving behaviour of three pelagic seabird species (*Gulosus aristotelis*, *Uria aalge* and *Alca torda*) from GPS data *alone* with greater accuracy than both hidden Markov models and a naive Bayes classifier (Browning et al., 2018). These findings not only attest to the untapped potential of DNNs as behavioural classification tools, but also importantly show how seabird foraging can be successfully monitored without the need for costly and cumbersome TDR devices. However, while this is certainly a step in the right direction, short battery life and behaviour impairment resulting from heavy device weight remain inhibitory for those who wish to adopt a GPS-only strategy. As such, there is still a strong incentive to investigate whether similar insights can be derived from a less obstructive and

longer-lasting bio-logging technology.

To this end, I here train a deep neural network using data labelled with depth-validated diving events to evaluate the utility of GLS-derived salt-water immersion data (IMM) as predictors of diving behaviour in pelagic seabirds, thereby providing the first known assessment of the potential for using GLS alone to identify foraging locations for wide-ranging species throughout their annual cycles. As a performance benchmark, I also assess the predictive power of high-resolution, tri-axial acceleration data (ACC) for the same task. For both predictors I address the additional question of how the width of the time window of data to be classified affects model performance. These methods are used here to investigate the foraging behaviour of the Red-Footed Booby (hereafter 'RFB') — a small, pantropical, diurnal, central-place foraging seabird widely understood to be the most pelagic of its genus (Nelson and Nelson, 1978, Schreiber et al., 1996) — in the BIOT MPA, Indian Ocean. It is initially hypothesized that the comparatively coarse IMM data will struggle to identify localised dive events but succeed in roughly capturing the broader areas in which they occur due to the species-specific changes in the birds' interactions with the water during active foraging bouts. Much stronger performances are expected in the case of ACC data, and it is predicted that these results will serve to inform long-term telemetric deployments in the future. It is also expected that smaller window widths will achieve greater classification accuracy scores for ACC data, owing to its granularity, and that the reverse will be true for IMM data.

2 Methods and Materials

2.1 Data

The British Indian Ocean Territory Marine Protected Area (BIOT MPA) is a contiguous, no-take marine reserve established by the British government that contains within it a Ramsar site and several Important Bird Areas (IBA), making it somewhat of a refuge for the 18 species of breeding seabird that live within it. The protected status of this area, along with the centralised foraging strategy employed by its resident RFBs when breeding, makes these birds suitable candidates for behavioural studies that require the deployment and eventual recovery of multiple logging devices, as it permits the observation of natural behaviour over the open ocean while remaining close enough to facilitate easy monitoring and access (Einoder, 2009).

The data reviewed in this project were collected from 9 breeding RFBs tracked in the BIOT MPA over a 3-6 day period in February 2019. All individuals were caught on Diego Garcia, the largest island of the Chagos Archipelago (7°18'48.0"S, 72°24'40.0"E), and fitted with high-sampling-rate multi-sensor GPS loggers (AxyTrek Marine, Technosmart, 14g, 1.4% bird body mass) and light-level geolocator/salt-water immersion loggers (Intigeo C330, Migrate Technology, 3g, 0.3% bird body mass). GPS loggers were configured to record geographical location every 30s, pressure every 1s, and tri-axial acceleration (ACC) at 25Hz. Geolocators,

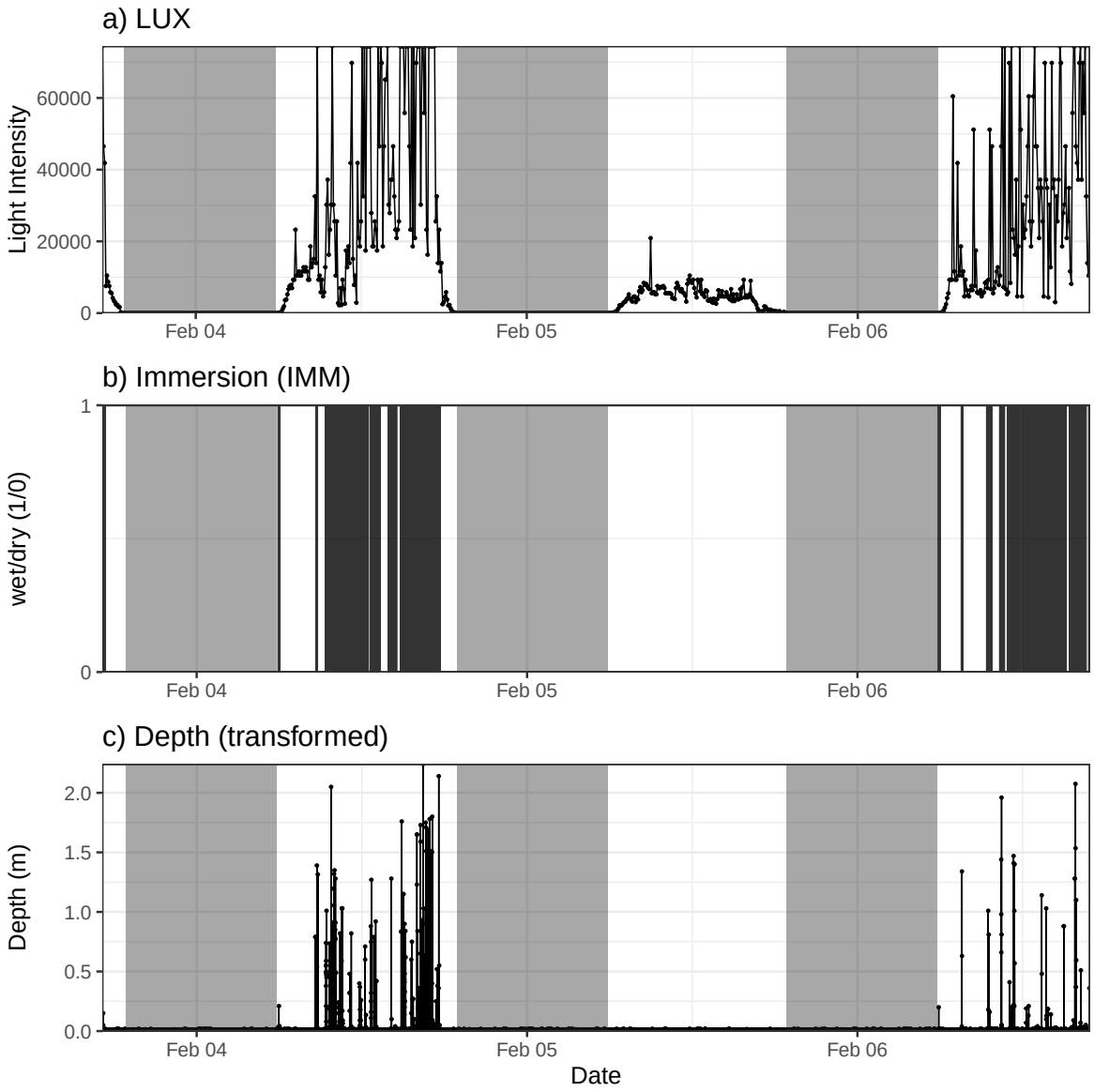


Figure 1: Example distribution of **a)** ambient light level data; **b)** binary salt-water immersion data; and **c)** de-noised depth data for a single bird (*ch_gps16_S1*) over the period it was tracked. Dark (night-time) periods are shown shaded grey. The plot shows how the bird interacts with the water exclusively during daylight hours and spends the whole of Feb 05 off the water, likely staying in a relatively shaded area around the colony. Spikes in depth (i.e. diving events) correlate strongly with periods immersed.

on the other hand, sampled at a much lower resolution, recording a binary immersion value (wet/dry) every 6s and the duration of time spent in each sequential state, as well as the light level at each minute, recording only the maximum value sampled in every 5 minute window. Figure 1 illustrates an example of how these data are distributed over the study period for a single bird. Of the 15 sets deployed, failure to recover hardware or subsequent data corruption meant that data could only be successfully extracted from 9 GPS and 6 GLS loggers, resulting in the availability of full data for 5 birds and GPS data for an additional 4.

2.2 Data Preprocessing

All preprocessing was carried out using the programming languages R (v4.1.0) and Python (v3.9.6). Firstly, raw pressure readings from the GPS logger were converted to depth data using the XManager software (`https://www.technosmart.eu/`), and subsequently 'de-noised' to control for device idiosyncrasies. This was achieved by offsetting all values within a rolling window of 30 records by the median of that window, effectively smoothing the low-level background fluctuations in each time-series whilst preserving the shape and location of each spike, thus enabling the determination of a fixed threshold for identifying dive events. For each bird, all data preceding the first aerial departure and following the final return to the island were clipped to ensure that none of the data reviewed here were influenced by device attachment/detachment processes (as multiple devices were logging data before and after their time fixed to a bird). GPS coordinates were also linearly interpolated in gaps exceeding 60s (i.e. where one or more readings had been missed). GLS files were then extracted and expanded to 6s resolution before being matched by ID to their corresponding GPS files (dropping excess data according to the same time frames) in order to append the transformed depth data within the latter to enable the identification of dive events by the same method. All in all, this resulted in a preliminary data set comprising 87,787,041 observations across 9 birds for ACC and 214,838 across 5 birds for GLS. To visualise these data, I mapped GPS tracks for all 9 birds over the MPA and calculated for each bird the total time tracked, total distance travelled (km), furthest distance travelled from the nest (assumed to be the mode lat/lon coordinates), maximum depth recorded, and the total number of observed dive and non-dive events (method for determining these described below).

From the processed data, several labelled data sets were then constructed for each predictor using a rolling window of raw data points as inputs and binary values indicating whether a dive had occurred within each given window as the corresponding classes. Dives were classified in windows where at least one depth value exceeded a predetermined threshold of 0.1m, which was chosen through visual inspection of depth time-series plots to distinguish periods of residual background noise from the periods of sharp fluctuation that are characteristic of genuine diving events. A range of window widths were tested for each predictor to determine the optimum. These were chosen to span time intervals that could reasonably be expected to capture dive behaviour based on the resolution and distribution of the data (2s, 4s, 6s, 8s, & 10s for ACC, and 1min, 3min, 5min, 7min, and 9min for IMM). To investigate the effects of diluting the ACC data on its predictive power, an additional data set was created using as inputs only the mean and sum of absolute differences of 4min windows of data points along the z-axis.

The method described above produced a number of large data sets that were highly imbalanced (up to 1TB for larger window widths, in which only $\sim$0.18% of rows contained dive behaviour), so to sidestep consequent performance and memory-related issues, a smaller, more balanced subset was extracted from each by randomly undersampling the majority class (i.e. non-dives) to roughly match the number of dive rows, reducing the maximum file

size to ∼2GB.

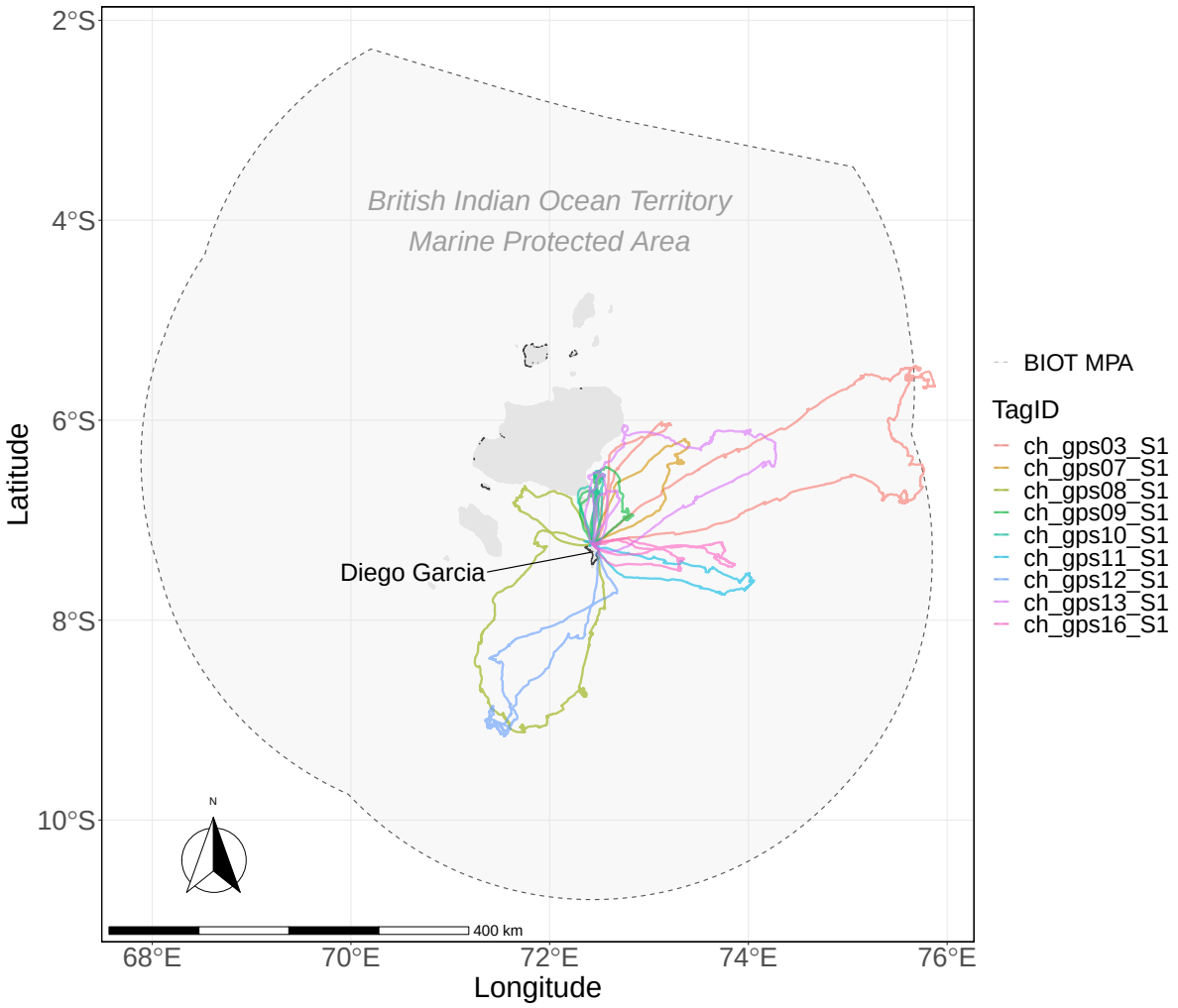
2.3 Model Fitting

Neural networks were used to predict behavioural states. Model fitting was performed primarily in Keras (v2.5.0) — a high-level API for interfacing with Google's Tensorflow library to build, train, and evaluate neural networks — using a basic architecture of 2 hidden layers of 200 rectifier nodes (each with a dropout rate of 0.2) followed by a softmax binary output layer. The shape of the input layer varied with the window width and number/resolution of variables used to generate the input data. IMM constitutes a single variable and was therefore associated with just one sequence of values per input, whereas a single ACC input consisted of just 2 summary statistics in the case of the diluted data and a concatenated vector of all 3 axes of the raw ACC data otherwise. Therefore, the size of the input layer for ACC ranged from 2 nodes where summary statistics were used to 750 where a width of 10s was used for the full, 25Hz tri-axial data, and, for IMM, from 10 nodes where a window width of 1min was used to 90 where a width of 9mins was used. A customised 'leave-one-out' cross-validation method was adopted whereby each model was validated with data from a withheld bird, ensuring generalisability of results and reflecting how such models would likely be used in practise. All models were saved for downstream analysis, resulting in a single saved model for each bird for each window width. To diagnose the performance of the IMM classifier, a separate classifier was constructed to simply predict dive events wherever windows contained a mix of wet and dry samples and non-dive events otherwise (i.e. where windows were all dry or all wet). This naive model was tested on the same, reduced data as the neural network to ensure valid performance comparison.

The classification accuracy, Area Under the ROC Curve (AUC), precision, sensitivity, and specificity were calculated for all models, as well as the scaled confusion matrix. These metrics were then averaged to produce single, cross-validated scores for each window width, which were used to determine the optimal for each predictor.

2.4 Georeferencing Predictions

The predictions of the best models were georeferenced to gain a deeper understanding of how these can be used to pinpoint foraging locations in practise. To this end, the timestamp at the midpoint of each input window was saved along with its corresponding class during the construction of the rolling window data to enable mapping to nearest GPS coordinates. GPS data were left raw for mapping high-resolution ACC predictions, but for the IMM predictions, to account for the relatively large window widths used to compensate for the data's low temporal resolution, GPS data were first collapsed to 5min resolution to ensure that the window surrounding each GPS coordinate contained within it the majority of several IMM windows rather than just a minority of a handful. (To be specific, keeping GPS data at 30s resolution would mean having the midpoint of 6 IMM windows falling within 15s of — and consequently being mapped to — each GPS coordinate, despite the fact that each of



TagID	Time Tracked (days)	Total Distance Travelled (km)	Max Distance Travelled from Colony (km)	Max Depth (m)	Dives	Non−dives
ch_gps03_S1	5.78	1969.23	420.19	3.89	162	16555
ch_gps07_S1	3.91	619.05	155.78	2.15	25	11300
ch_gps08_S1	5.24	1268.27	226.03	2.58	154	14979
ch_gps09_S1	4.32	790.10	86.34	1.91	71	12397
ch_gps10_S1	4.31	596.70	81.38	1.90	34	12438
ch_gps11_S1	3.19	731.33	182.49	2.59	67	9149
ch_gps12_S1	4.10	1027.90	234.98	2.39	52	11807
ch_gps13_S1	5.78	1308.28	231.10	2.60	109	16582
ch_gps16_S1	4.02	902.01	157.68	2.53	135	11491

Figure 2: Flight paths for all 9 GPS-tagged birds in the BIOT MPA (outlined). Summary statistics for each bird are also shown, with birds from whom GLS data were also successfully recovered highlighted in blue. (One outlier was removed where the TDR fixed to bird *ch_gps03_S1* recorded a depth value of over 10m.)

Predictor	Window Width	Accuracy	AUC	Precision	Sensitivity	Specificity
ACC	2s	98.42%	98.96%	98.86%	97.89%	98.94%
	4s	98.5%	98.91%	98.62%	98.33%	98.71%
	6s	98.17%	98.53%	98.43%	97.85%	98.44%
	8s	98.24%	98.53%	97.92%	98.51%	97.96%
	10s	97.45%	97.79%	98.05%	96.71%	98.04%
IMM	1min	81.41%	85.15%	92.02%	67.63%	93.96%
	3min	92.81%	95.7%	91.42%	94.06%	91.22%
	5min	93.67%	96.96%	91.04%	96.62%	90.71%
	7min	93.06%	97.09%	89.36%	97.61%	88.43%
	9min	92.66%	97.07%	89.96%	96.45%	88.92%

Figure 3: Table of cross-validated, mean classification metrics for each window width tested for each predictor. Using accuracy as a guide, optimal predictions were produced when a 4s window width was used for ACC data and a 5min window width was used for IMM data.

these windows may cover up to 18 other GPS coordinates also!) This way, the class of a GPS location could be ascertained by examining the proportion of dive predictions in its immediate vicinity. If this proportion exceeded a certain threshold, then the GPS coordinate would qualify as a dive location. This method of georeferencing was applied for both predictors, with the threshold set to a default value of 50% in each case (the effect of changing this value was not investigated here, but could be subject to further inquiry - see Discussion).

3 Results

Figure 2 shows the mapped GPS tracks for all 9 birds and accompanying summary statistics. Classification metrics for all predictive models are summarised in Figure 3.

3.1 ACC data

Using raw, tri-axial ACC to predict dive events, it was found that all leave-one-out cross-validated models scored mean accuracy values of above 97%, demonstrating strong predictive power for this data that is robust to a withheld bird. Additionally, sensitivity and specificity remained high across all window widths tested, indicating a general proficiency at detecting both dive events and non-dive events. The optimal model was produced where a 4s time window was used to generate training data (see Fig. 4 for the resulting confusion matrix, and Fig. 5 for the distribution of metrics across cross-validation folds for this model), scoring a mean accuracy of 98.5% that varied negligibly across folds of the cross-validation

procedure (Fig. 5). Mean sensitivity and precision scores for this model were each second highest of those tested, representing the best combination at minimising the number of genuine dive sites missed (false negatives) while preventing overclassification by penalising bogus dive sites (false positives).

Models trained only with z-axis ACC summary statistics performed substantially worse, scoring a mean classification accuracy of 87.15% (sensitivity=95.6%, specificity=78.7%), indicating that diluting the data in this way preserves some measure of what constitutes a dive but also drives the misclassification of non-dive events.

3.2 IMM data

The results of models using IMM data to predict dive events showed increased variation and slightly reduced accuracy, ranging from 81.41% to 93.67% where window widths of 1min and 5min were used respectively (Fig. 3). Precision saw the most significant reductions of all metrics compared to ACC data, dropping by a mean average of nearly 8% (Fig. 5). While sensitivity scaled roughly in proportion to window width, a negative correlation was observed between window width and precision (and also window width and specificity), suggesting that larger window widths capture higher proportions of genuine dive events but also drive the misclassification of non-dive events for this predictor. The optimal window width was accordingly the median of those tested (see Fig. 4 for the resulting confusion matrix, and Fig. 5 for the distribution of metrics across cross-validation folds for this model). When cross-validating the predictions of this model, the model evaluated using data from unseen bird *ch_gps12_S1* scored significantly lower precision (85.92%) than the others (91.42%+),

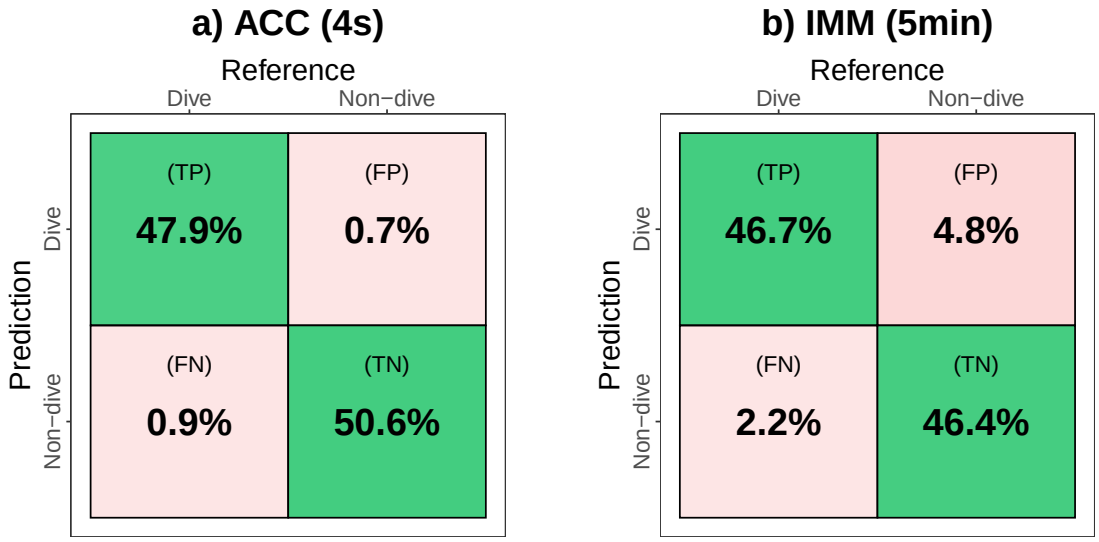


Figure 4: Scaled confusion matrices for the optimal models of both **a)** ACC; and **b)** IMM data, showing the proportion of total data points in each classification bucket.

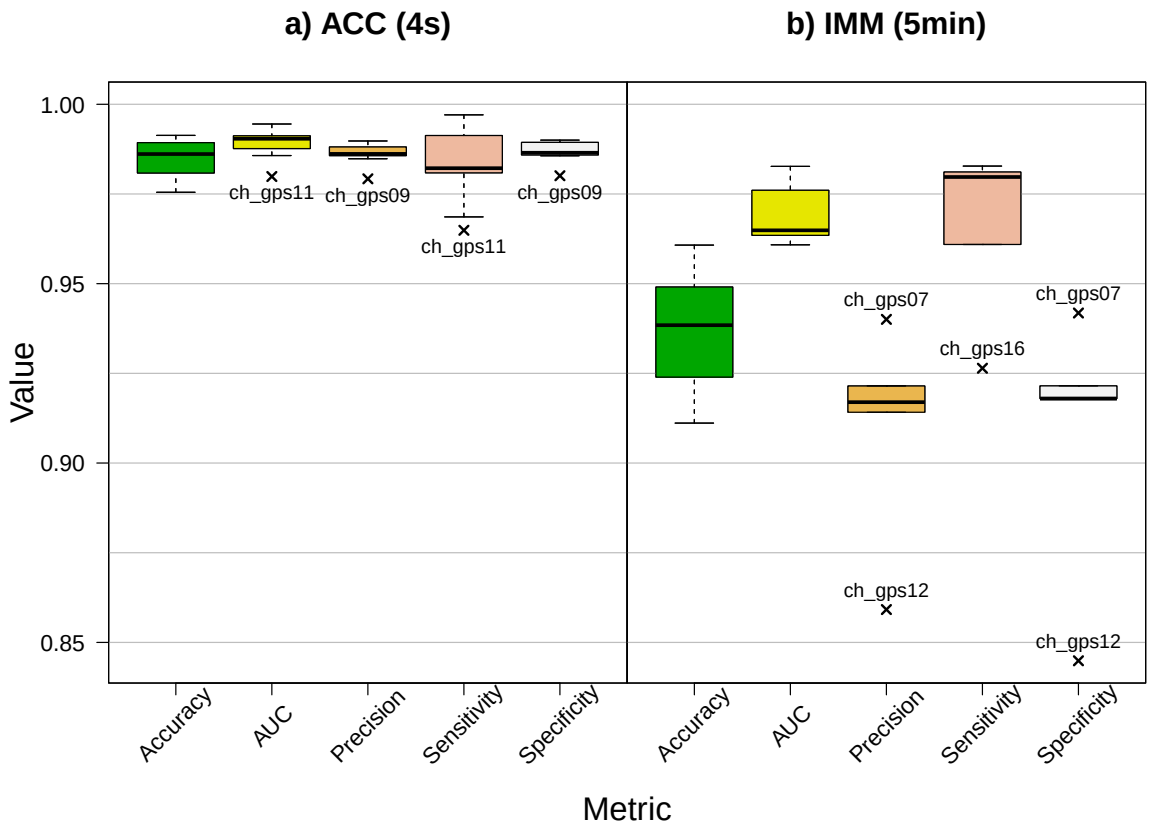


Figure 5: Box plots showing the distribution of classification metrics scored by optimal models for both predictors across folds of the leave-one-out cross-validation process. Each point therefore corresponds to a bird withheld for model testing. Outliers are labelled with tag ID.

suggesting that this bird often exhibited 'dive-looking' behaviour while not actually foraging (Fig. 4).

Interestingly, the naive IMM classifier correctly classified 92.77% of the same data. Sensitivity remained high for this model at 97.62%, meaning that diving events were generally identified correctly, whereas specificity and precision fell to 88.13% and 88.71% respectively, indicating that the increase in sensitivity is likely due to an overclassification of diving events and that the marginal performance gap between the naive classifier and the DNN is due to the latter's superior ability to discriminate windows containing a mixture of wet and dry samples.

3.3 Georeferencing Predictions

When mapped to nearest GPS coordinates, the distribution of dive and non-dive predictions returned to an imbalance comparable to that of the data prior to subsetting (Fig. 6). This is due to the fact that the majority of dive samples produced in the rolling window data subsets were generated around single, highly localised events, meaning that any positive predictions corresponding to these samples collapsed back into a small handful of points

when georeferenced. Both predictors remained highly proficient at detecting genuine dive and non-dive sites (sensitivity=96.43% and specificity=99.16% for ACC; sensitivity=96.55% and specificity=95.23% for IMM - calculated from the confusion matrix statistics displayed in Figure 6). Naturally, however, the increase in non-dive sites pushed up the relative incidence rate of true negatives and false positives. As a consequence, the negative predictive value of this georeferencing method was near-perfect (99.96% for ACC, 99.81% for IMM), while the precision was significantly lower (56.25% for ACC, 51.85% for IMM). In short, this means that this method is excellent at identifying both dive and non-dive events, misidentifying only a marginal proportion for each predictor, but the abundance of non-dive samples makes these proportions amount to substantially large quantities compared to the relatively rare incidence of dives, bumping the incidence of false positives up to near that of true positives in both cases.

4 Discussion

Using the combined data from GLS and integrated GPS/TDR devices, I trained a deep neural network to compare the performances of 25Hz tri-axial acceleration data (ACC) and 1/6Hz salt-water immersion data (IMM) as predictors of diving behaviour of Red Footed Boobies. Predictions were validated with diving events identified using depth data. It was found that DNNs can predict the behavioural status (dive/non-dive) of a 4s window of 25Hz tri-axial ACC data with 98.5% accuracy and the status of a 5min window of 1/6Hz IMM data with 93.67% accuracy, and that IMM data could correctly classify similar proportions of genuine diving events to ACC data but significantly less non-diving events. From these findings it seems plausible that diving behaviours, and, by proxy, foraging locations, can be mapped for wide-ranging animals throughout their annual cycles while keeping device weight (and its consequent behavioural confounds) to a minimum; a promising result in a field where such insights have hitherto been available only for for larger birds over short time-frames. There are, however, some noteworthy caveats to doing this in practise.

Firstly, it is important to consider the way the depth threshold used to identify dive events is established for each bird when labelling the training data. Here, once the depth distributions had been ‘de-noised’, a fixed threshold of 0.1m was determined by eye that subsequently applied for all birds. But while it is true that this threshold seems to have successfully distinguished diving behaviour for the most part, it is also possible that idiosyncrasies in both individual bird behaviour and device sampling could make a single threshold inappropriate for multiple birds. For studies wishing to replicate the methods described here, a more statistically rigorous method of establishing this value, or multiple values for multiple birds (e.g. using change-point analysis), would likely improve the labelling accuracy of the data and consequently reduce the errors made by models trained with it.

The foraging behaviours of the particular bird species under review have important implications for data labelling also. Red-footed boobies are shallow-diving seabirds that

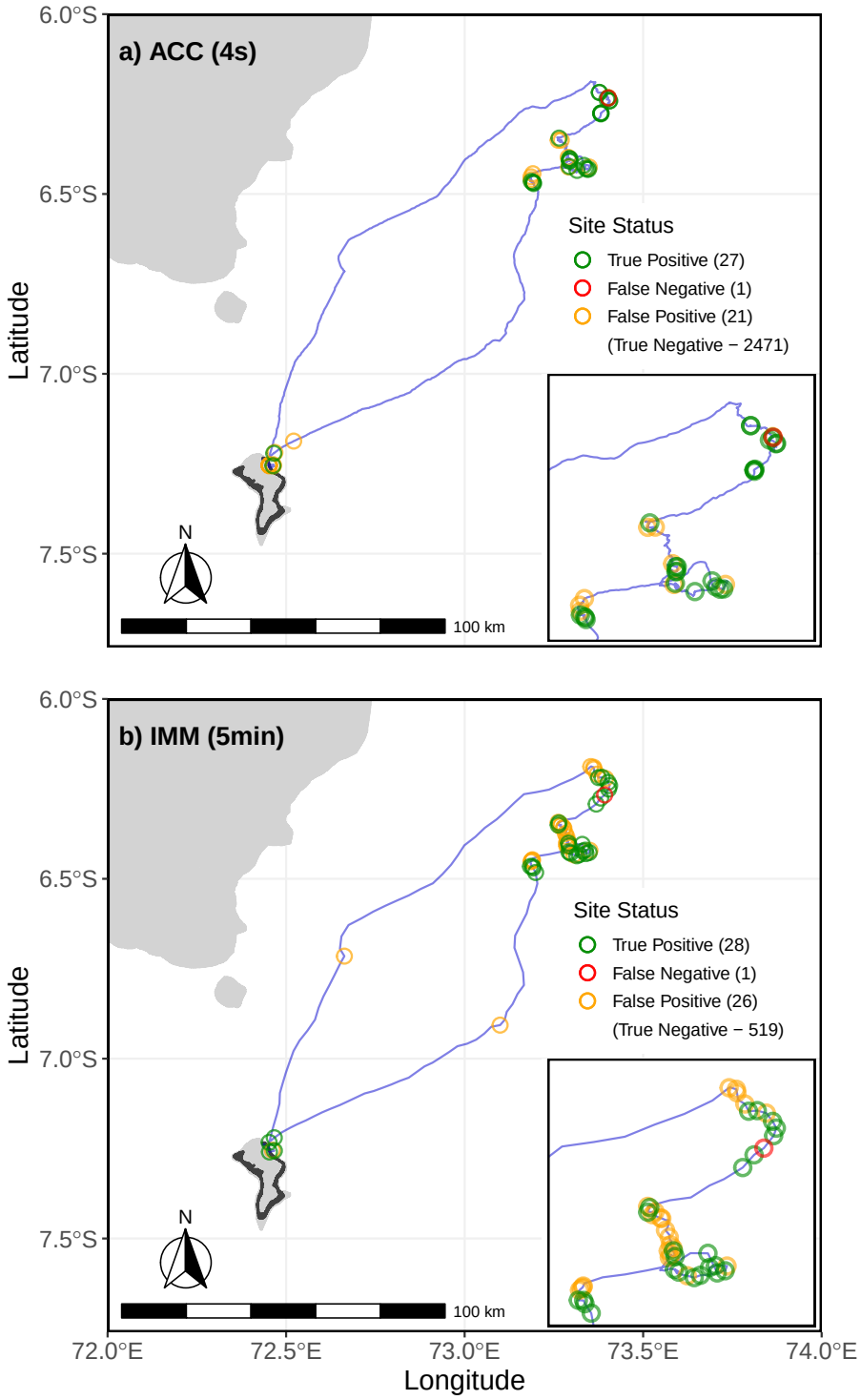


Figure 6: Example of a GPS track with georeferenced model predictions overlaid for **a)** 4s windows of ACC data; and **b)** 5min windows of IMM data from the reduced data for bird *ch_gps07_S1*. True negatives are included in the legend but omitted from the track for visual clarity. Points are partially transparent, meaning that darker hues indicate clusters of overlapping points in the same region. Mismatch of total no. of predictions between the two maps is due to the differing resolutions of the GPS data and the predictors between them. Tracks show an increase in the proportion of false positives compared to the distribution of errors in the predictions prior to mapping, yet largely clustered around true positives and sparse elsewhere.

engage in a variety of diving behaviours when foraging, including surface dives, plunge dives from low altitudes, and, less frequently, aerial dives to catch aerially locomotive prey such as flying fish or squid above the surface (Weimerskirch et al., 2005, Diamond, 1974, Pitman and Ballance, 1993). The mix of shallow diving behaviours and near-surface non-foraging activity complicates inferring any distinctions from depth data alone, as, for example, the depth value separating a shallow surface dive from a period merely sitting on the surface with legs submerged may be obscure. (In the case of IMM data for RFBs in particular, aerial foraging activity will be missed altogether!) Thus, for species whose interactions with the water during foraging are more distinctive, such as deep-divers or those who engage in an active swim phase (e.g. Gannets), determining labels in the training data should be less problematic, leading to all-round better model performance.

Another important set of considerations for improving model performance are to do with the later step of model training. It is important to note, for instance, that all models tested here were trained on reduced data sets, and as such did not see the vast majority of data pertaining to each bird. Therefore, when applied to a full data set for georeferencing, it is likely that the sharp increase in negative (i.e non-dive) samples to be classified will drastically ramp up the incidence of false positives. As a potential solution, it should be remembered that the data sets were reduced in the first place to rectify the considerable class imbalance by undersampling the majority class (non-dives), but there are various techniques for doing the reverse and oversampling the minority class if performance requires it. For example, one method (dubbed 'SMOTE' by its inventors: Chawla et al. (2002)) involves generating synthetic samples using a k-nearest neighbours approach to augment the minority class and balance the class distribution, thereby enabling the model to see all available data and still learn the decision boundary effectively. As all models evaluated in this project were only exposed to a tiny minority of the available 'non-dive' data during training, it is possible that important insights as to what constitutes a 'non-dive' sample were missed, and accordingly that model precision (and, indeed, overall performance) could have been further improved by adopting such a method as SMOTE for balancing the class distribution before training.

It could also be that the observed spike in false positives can be reduced further, in quantity or significance, in the final step step of georeferencing predictions. An important observation from Figure 6 is that false positives seem to generally cluster around true positives and are spread much more sporadically elsewhere. Accordingly, if the aim is to roughly identify regions in which foraging behaviours occur, then it could be argued that the occurrence of the former when implementing this georeferencing method is benign. The more isolated false positives, however, pose more of a problem as they have the potential to lead mapping efforts further off the mark than is acceptable. But it is possible that these can be resolved simply by including more data and/or adjusting the threshold used to classify individual sites. This is because while many of the isolated false positives may be legitimate, the fact that only random subsets of non-dives were included in the data used here means that some of the non-dive sites that were misidentified as dive sites by the model may only have had a single prediction in their vicinity, driving the proportion of local

dive predictions to 1 and classifying that location as a dive location, even though that same prediction could possibly have just been one of a small minority of predictions in the same window had the whole data set been used. Therefore, so long as a suitable threshold can be determined for distinguishing clusters of positive predictions from solitary ones, then it is likely that many of the more isolated false positives observed here would disappear, thereby refining model precision and minimising the erroneous highlighting of insignificant areas.

4.1 ACC to predict foraging behaviour

Irrespective of future use of IMM data for the means described here, it can be safely concluded from my results that ACC data is an excellent predictor of diving behaviour in pelagic seabirds. While it is true that this revelation will not be so useful to those wishing to map foraging locations for birds for whom GLS data has already been collected, it certainly attests to the potential merits of using accelerometers alongside GLS in future studies aiming to do something similar. For this to be feasible, an accelerometer would need to weigh no more than a few grams and have a long operational life (preferably over a year), yet store enough data at a high enough resolution to capture the information required to identify fleeting behaviours that occur infrequently. As acknowledged by Bäckman et al. (2017) in their review of the usage of accelerometers in migration studies, the development of a device that meets all these requirements is a challenge, but has been successfully achieved in a number of studies. Liechti et al. (2013), for example, used a 1.5g data logger configured to record light intensity and dynamic acceleration every 4mins by sampling 32 measurements of 10Hz acceleration along the z-axis — recording only the mean of each sample and sum of the absolute differences between each consecutive data point in each sample — that lasted year-round, and Hedenström et al. (2016) used a similar approach to record acceleration and light level data over a full 2 year period.

It remains unclear precisely what effect such resolution cutbacks would have on the performance of a model built to predict diving behaviour specifically using ACC data. The preliminary results of the models trained here using only z-axis summary statistics suggest that predictive accuracy would suffer. But it is important to note firstly that these cutbacks are only required inasmuch as longevity is necessary to any given study, and secondly that, for those to which it *is* necessary, it is highly possible that enough information can be captured by the right summary statistics (or perhaps their combination with some other predictor) to accurately identify diving behavior nonetheless. Accordingly, while the advent of lightweight geolocators with inbuilt accelerometers leaves open the possibility that the predictive power of ACC data demonstrated here can still be leveraged in more longitudinal studies of wider-ranging birds, it cannot be commented on with any certainty until the performance tradeoffs of diluting ACC data in the ways mentioned above are investigated further.

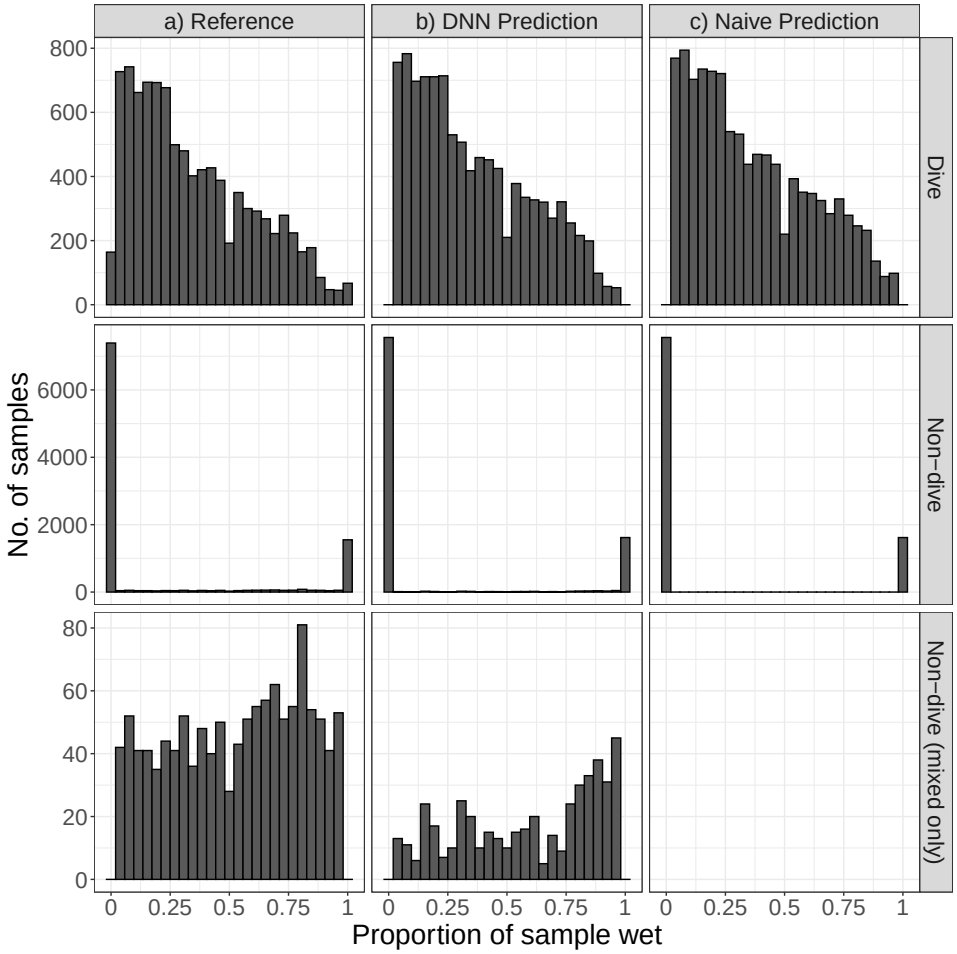


Figure 7: Facet-wrapped histogram showing the disribution of wet/dry samples across all 5min windows used in the reduced IMM dataset, and how they were classfied by: **a)** their corresponding depth-validated labels; **b)** the neural network; and **c)** the naive classifier. It is clear from the top two rows that both the DNN and naive model classify all homogeneous windows (i.e. all wet or all dry) as non-dives. The bottom row shows that the DNN better discriminates windows containing a mix of wet and dry samples.

4.2 IMM to predict diving behaviour

On the other hand, the results described in this project provide good reason to believe that diving behaviours, and, by proxy, foraging locations, can also be predicted for animals for whom only GLS data are available. Accordingly, it seems plausible that the methods described here can be applied both in future studies assessing the long-term behavioural patterns of wide-ranging birds (using GLS either singularly or in combination with one or more other informative, long-term predictors, such as ACC) *and* to the archive of GLS data already available from past projects.

It must be remembered, though, that the geographic accuracy of mapped predictions will naturally scale to the spatio-temporal resolution of the tracking method used. As geolocation is considerably less precise than GPS, the time window surrounding each set of derived coordinates would be larger. Consequently, one would ultimately be required to collapse

significantly more IMM predictions into each location when georeferencing them, leading to spatial regions far more coarse than those observed here, but containing more granular (and possibly more informative) distributions of predictions within them. Therefore, it is possible that spatial resolution could give way to more reliable predictions when using geolocation rather than GPS to map them.

The strong predictive accuracy achieved by the naive IMM classifier points to some interesting considerations regarding species-specific foraging behaviours and their implications for model classification heuristics also. Relevant to this particular project, for instance, is the knowledge that RFBs land on the sea surface much more frequently than most other seabirds during active foraging bouts (Weimerskirch et al., 2005). It accordingly makes sense that the naive model should be accurate, as it could be speculated from this premise *a priori* that most windows containing a mix of wet and dry samples will either contain, or fall close to, a genuine dive event; a hypothesis confirmed by subsequent analysis of the distribution of wet/dry samples across all predicted data windows (Fig. 7). (It follows that the value of the DNN here lies in its ability to better discriminate the attributes of the mixed samples. However, the opaque nature of the technique precludes further analysis of how this is achieved.) Similarly, it is not difficult to imagine that an analogous — and perhaps even more effective — heuristic for identifying dive events from IMM data could exist for other species whose interactions with the water are almost exclusively linked to foraging, obviating the need for a sophisticated model altogether in cases where quick approximations are needed.

However, the converse is also true: immersion patterns of species with more complex interactions with the sea surface may be too difficult to accurately classify, even for a DNN, and while it is true that model performance could probably be improved with the inclusion of other GLS products such as LUX data (e.g. by assigning lower probabilities to 'foraging-like' behaviours observed at low light levels for diurnal species, thereby accounting for any temporal element of the animal's feeding cycle), this obstacle may prove insurmountable without more granular data. Therefore, while the ability of immersion data to predict diving behaviour demonstrated here is certainly promising, it cannot necessarily be generalised to birds in other foraging guilds without first taking such considerations into account.

Data and Code Availability

- Data:
  `https://dropbox.com/sh/llxrfoi53yidvdt/AAA4b1NxKdDhuXNpjNOZzUs-a?dl=0`
- Code:
  `https://github.com/ldswaby/seabirds`

Please note that in order for the code to execute successfully the Data/ directory must be in the same directory as the Code/ directory. The required directory structure can be viewed in the README.md file in the github repository.

References

Åkesson, S., Klaassen, R., Holmgren, J., Fox, J. W. and Hedenström, A. (2012), 'Migration routes and strategies in a highly aerial migrant, the common swift apus apus, revealed by light-level geolocators', *PloS one* **7**(7), e41195.

Bächler, E., Hahn, S., Schaub, M., Arlettaz, R., Jenni, L., Fox, J. W., Afanasyev, V. and Liechti, F. (2010), 'Year-round tracking of small trans-saharan migrants using light-level geolocators', *PloS one* **5**(3), e9566.

Bäckman, J., Andersson, A., Pedersen, L., Sjöberg, S., Tøttrup, A. P. and Alerstam, T. (2017), 'Actogram analysis of free-flying migratory birds: new perspectives based on acceleration logging', *Journal of Comparative Physiology A* **203**(6-7), 543–564.

Barron, D. G., Brawn, J. D. and Weatherhead, P. J. (2010), 'Meta-analysis of transmitter effects on avian behaviour and ecology', *Methods in Ecology and Evolution* **1**(2), 180–187.

Bishop, C. M. et al. (1995), *Neural networks for pattern recognition*, Oxford university press.

Bograd, S. J., Block, B. A., Costa, D. P. and Godley, B. J. (2010), 'Biologging technologies: new tools for conservation. introduction', *Endangered Species Research* **10**, 1–7.

Bost, C. and Le Maho, Y. (1993), 'Seabirds as bio-indicators of changing marine ecosystems: new perspectives', *Acta Oecologica-International Journal of Ecology* **14**(3), 463–470.

Breed, G. A., Costa, D. P., Jonsen, I. D., Robinson, P. W. and Mills-Flemming, J. (2012), 'State-space methods for more completely capturing behavioral dynamics from animal tracks', *Ecological Modelling* **235**, 49–58.

Browning, E., Bolton, M., Owen, E., Shoji, A., Guilford, T. and Freeman, R. (2018), 'Predicting animal behaviour using deep learning: Gps data alone accurately predict diving in seabirds', *Methods in Ecology and Evolution* **9**(3), 681–692.

Calvo, B. and Furness, R. (1992), 'A review of the use and the effects of marks and devices on birds', *Ringing & Migration* **13**(3), 129–151.

Carignan, V. and Villard, M.-A. (2002), 'Selecting indicator species to monitor ecological integrity: a review', *Environmental monitoring and assessment* **78**(1), 45–61.

Carroll, G., Slip, D., Jonsen, I. and Harcourt, R. (2014), 'Supervised accelerometry analysis can identify prey capture by penguins at sea', *Journal of Experimental Biology* **217**(24), 4295–4302.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002), 'Smote: synthetic minority over-sampling technique', *Journal of artificial intelligence research* **16**, 321–357.

Ciancio, J. E., Yorio, P., Buratti, C., Colombo, G. Á. and Frere, E. (2021), 'Isotopic niche plasticity in a marine top predator as indicator of a large marine ecosystem food web status', *Ecological Indicators* **126**, 107687.

Dean, B., Freeman, R., Kirk, H., Leonard, K., Phillips, R. A., Perrins, C. M. and Guilford, T. (2013), 'Behavioural mapping of a pelagic seabird: combining multiple sensors and a hidden markov model reveals the distribution of at-sea behaviour', *Journal of the Royal Society Interface* **10**(78), 20120570.

Diamond, A. (1974), 'The red-footed booby on aldabra atoll, indian ocean', *Ardea* **62**(3-4), 196–218.

Einoder, L. D. (2009), 'A review of the use of seabirds as indicators in fisheries and ecosystem management', *Fisheries Research* **95**(1), 6–13.

Gillies, N., Fayet, A. L., Padget, O., Syposz, M., Wynn, J., Bond, S., Evry, J., Kirk, H., Shoji, A., Dean, B. et al. (2020), 'Short-term behavioural impact contrasts with long-term fitness consequences of biologging in a long-lived seabird', *Scientific reports* **10**(1), 1–10.

Grünewälder, S., Broekhuis, F., Macdonald, D. W., Wilson, A. M., McNutt, J. W., Shawe-Taylor, J. and Hailes, S. (2012), 'Movement activity based classification of animal behaviour with an application to data from cheetah (acinonyx jubatus)', *PloS one* **7**(11), e49120.

Guilford, T., Meade, J., Freeman, R., Biro, D., Evans, T., Bonadonna, F., Boyle, D., Roberts, S. and Perrins, C. (2008), 'Gps tracking of the foraging movements of manx shearwaters puffinus puffinus breeding on skomer island, wales', *Ibis* **150**(3), 462–473.

Guilford, T., Meade, J., Willis, J., Phillips, R. A., Boyle, D., Roberts, S., Collett, M., Freeman, R. and Perrins, C. (2009), 'Migration and stopover in a small pelagic seabird, the manx shearwater puffinus puffinus: insights from machine learning', *Proceedings of the Royal Society B: Biological Sciences* **276**(1660), 1215–1223.

Hazen, E. L., Abrahms, B., Brodie, S., Carroll, G., Jacox, M. G., Savoca, M. S., Scales, K. L., Sydeman, W. J. and Bograd, S. J. (2019), 'Marine top predators as climate and ecosystem sentinels', *Frontiers in Ecology and the Environment* **17**(10), 565–574.

Hedenström, A., Norevik, G., Warfvinge, K., Andersson, A., Bäckman, J. and Åkesson, S. (2016), 'Annual 10-month aerial life phase in the common swift apus apus', *Current Biology* **26**(22), 3066–3070.

Jackson, S. and Wilson, R. P. (2002), 'The potential costs of flipper-bands to penguins', *Functional Ecology* **16**(1), 141–148.

Jahn, A. E., Levey, D. J., Cueto, V. R., Ledezma, J. P., Tuero, D. T., Fox, J. W. and Masson, D. (2013), 'Long-distance bird migration within south america revealed by light-level geolocators', *The Auk* **130**(2), 223–229.

Jonsen, I. D., Flemming, J. M. and Myers, R. A. (2005), 'Robust state–space modeling of animal movement data', *Ecology* **86**(11), 2874–2880.

LeCun, Y., Bengio, Y. and Hinton, G. (2015), 'Deep learning', *nature* **521**(7553), 436–444.

Liechti, F., Witvliet, W., Weber, R. and Bächler, E. (2013), 'First evidence of a 200-day non-stop flight in a bird', *Nature Communications* **4**(1), 1–7.

Maclean, I. M., Rehfisch, M. M., Skov, H. and Thaxter, C. B. (2013), 'Evaluating the statistical power of detecting changes in the abundance of seabirds at sea', *Ibis* **155**(1), 113–126.

Mallory, M. L., Robinson, S. A., Hebert, C. E. and Forbes, M. R. (2010), 'Seabirds as indicators of aquatic ecosystem conditions: a case for gathering multiple proxies of seabird health', *Marine Pollution Bulletin* **60**(1), 7–12.

Martiskainen, P., Järvinen, M., Skön, J.-P., Tiirikainen, J., Kolehmainen, M. and Mononen, J. (2009), 'Cow behaviour pattern recognition using a three-dimensional accelerometer and support vector machines', *Applied animal behaviour science* **119**(1-2), 32–38.

Minton, C., Gosbell, K., Johns, P., Christie, M., Fox, J. W. and Afanasyev, V. (2010), 'Initial results from light level geolocator trials on ruddy turnstone arenaria interpres reveal unexpected migration route', *Wader Study Group Bulletin* **117**(1), 9–14.

Nathan, R., Spiegel, O., Fortmann-Roe, S., Harel, R., Wikelski, M. and Getz, W. M. (2012), 'Using tri-axial acceleration data to identify behavioral modes of free-ranging animals: general concepts and tools illustrated for griffon vultures', *Journal of Experimental Biology* **215**(6), 986–996.

Nelson, S. and Nelson, B. (1978), *The Sulidae: gannets and boobies*, number 154, Oxford University Press, USA.

Newton, I. (2010), *The migration ecology of birds*, Elsevier.

Parsons, M., Mitchell, I., Butler, A., Ratcliffe, N., Frederiksen, M., Foster, S. and Reid, J. B. (2008), 'Seabirds as indicators of the marine environment', *ICES Journal of Marine Science* **65**(8), 1520–1526.

Patterson, T. A., Basson, M., Bravington, M. V. and Gunn, J. S. (2009), 'Classifying movement behaviour in relation to environmental conditions using hidden markov models', *Journal of Animal Ecology* **78**(6), 1113–1123.

Phillips, R. A., Xavier, J. C. and Croxall, J. P. (2003), 'Effects of satellite transmitters on albatrosses and petrels', *The Auk* **120**(4), 1082–1090.

Phillips, R., Silk, J., Croxall, J., Afanasyev, V. and Briggs, D. (2004), 'Accuracy of geolocation estimates for flying seabirds', *Marine Ecology Progress Series* **266**, 265–272.

Pitman, R. and Ballance, L. (1993), Booby prey-capturing behavior in the eastern tropical pacific, *in* 'Abstract. Pacific Seabird Group Twentieth Annual Meeting. Seattle, Washington'.

Rutz, C. and Hays, G. C. (2009), 'New frontiers in biologging science'.

Schreiber, E. A., Schreiber, R. W. and Schenk, G. A. (1996), *Red-footed Booby: Sula Sula*, American Ornithologists' Union.

Shoji, A., Elliott, K., Fayet, A., Boyle, D., Perrins, C. and Guilford, T. (2015), 'Foraging behaviour of sympatric razorbills and puffins', *Marine Ecology Progress Series* **520**, 257–267.

Thaxter, C. B., Lascelles, B., Sugar, K., Cook, A. S., Roos, S., Bolton, M., Langston, R. H. and Burton, N. H. (2012), 'Seabird foraging ranges as a preliminary tool for identifying candidate marine protected areas', *Biological Conservation* **156**, 53–61.

Urbano, F., Cagnacci, F., Calenge, C., Dettki, H., Cameron, A. and Neteler, M. (2010), 'Wildlife tracking data management: a new vision', *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**(1550), 2177–2185.

Wanless, S., Harris, M. P., Burger, A. E. and Buckland, S. T. (1997), 'Use of time-at-depth recorders for estimating depth and diving performance of european shags (registradores del uso del tiempo-a-profundidad para estimar utilizacion de las profundidades y rendimiento de zambullida de phalacrocorax aristotelis)', *Journal of Field Ornithology* pp. 547–561.

Weimerskirch, H., Le Corre, M., Ropert-Coudert, Y., Kato, A. and Marsac, F. (2005), 'The three-dimensional flight of red-footed boobies: adaptations to foraging in a tropical environment?', *Proceedings of the Royal Society B: Biological Sciences* **272**(1558), 53–61.