Received: 18 April 2018 | Revised: 11 October 2018 | Accepted: 30 October 2018

DOI: 10.1002/ece3.4740

ORIGINAL RESEARCH

WILEY Ecology and Evolution Open Access

A comparison of techniques for classifying behavior from accelerometers for two species of seabird

Allison Patterson[1] | Hugh Grant Gilchrist[2] | Lorraine Chivers[3] | Scott Hatch[3] | Kyle Elliott[1]

[1]Department of Natural Resource Sciences, McGill University, Ste Anne-de-Bellevue, Quebec, Canada

[2]Environment and Climate Change Canada, National Wildlife Research Centre, Ottawa, Ontario, Canada

[3]Institute for Seabird Research and Conservation, Anchorage, Alaska

*Correspondence
Allison Patterson, Department of Natural Resource Sciences, McGill University, Ste Anne-de-Bellevue, QC, Canada.
Email: Allison.patterson@mail.mcgill.ca

Funding information
British Ornithologists' Union; Natural Sciences and Engineering Research Council of Canada; McGill University, Grant/Award Number: 2015-7599 and F11-020; University of Manitoba; Northern Contaminants Program - Indian and Northern Affairs Canada; Newfoundland and Labrador Murre Conservation Fund - Wildlife Habitat Canada and Bird Studies Canada

Abstract

The behavior of many wild animals remains a mystery, as it is difficult to quantify behavior of species that cannot be easily followed throughout their daily or seasonal movements. Accelerometers can solve some of these mysteries, as they collect activity data at a high temporal resolution (<1 s), can be relatively small (<1 g) so they minimally disrupt behavior, and are increasingly capable of recording data for long periods. Nonetheless, there is a need for increased validation of methods to classify animal behavior from accelerometers to promote widespread adoption of this technology in ecology. We assessed the accuracy of six different behavioral assignment methods for two species of seabird, thick-billed murres (*Uria lomvia*) and black-legged kittiwakes (*Rissa tridactyla*). We identified three behaviors using tri-axial accelerometers: standing, swimming, and flying, after classifying diving using a pressure sensor for murres. We evaluated six classification methods relative to independent classifications from concurrent GPS tracking data. We used four variables for classification: depth, wing beat frequency, pitch, and dynamic acceleration. Average accuracy for all methods was >98% for murres, and 89% and 93% for kittiwakes during incubation and chick rearing, respectively. Variable selection showed that classification accuracy did not improve with more than two (kittiwakes) or three (murres) variables. We conclude that simple methods of behavioral classification can be as accurate for classifying basic behaviors as more complex approaches, and that identifying suitable accelerometer metrics is more important than using a particular classification method when the objective is to develop a daily activity or energy budget. Highly accurate daily activity budgets can be generated from accelerometer data using multiple methods and a small number of accelerometer metrics; therefore, identifying a suitable behavioral classification method should not be a barrier to using accelerometers in studies of seabird behavior and ecology.

KEYWORDS
accelerometer, animal behavior, behavioral classification, movement ecology, *Rissa tridactyla*, seabird tracking, *Uria lomvia*

1 | INTRODUCTION

Developments in biologging technology have greatly advanced our ability to study wildlife throughout their ranges, without restrictions and bias imposed by human observation and accessibility (Cagnacci, Boitani, Powell, & Boyce, 2010; Hebblewhite & Haydon, 2010). Traditional methods for measuring animal activity involve direct observation of animals in the field, which is labor intensive. Direct observation limits the scale of observations to times and locations where focal species are accessible to biologists, and creates opportunity for bias if focal animals, or their predators and prey, change behavior in response to the presence of observers (MacArthur, Geist, & Johnston, 1982; Quiros, 2007). Measuring animal activity with accelerometers overcomes most of these challenges by continuously logging activity wherever the individual goes, and, if small enough, with very little impact on the animal's behavior (Wilmers et al., 2015). Accelerometers have been used to answer a wide-range of ecological questions relating to prey capture (Sato et al., 2015), energetics (Elliott, Chivers, et al., 2014; Robson, Chauvaud, Wilson, & Halsey, 2012), physiology (Watanuki, Niizuma, Geir, Sato, & Naito, 2003), migration strategies (Bishop et al., 2015; Wiemerskirch, Bishop, Jeanniard-du-Dot, Prudor, & Sachs, 2016); but perhaps the most widespread application of accelerometers is in obtaining time-activity budgets (Berlincourt, Angel, & Arnould, 2015; Brown, Kays, Wikelski, Wilson, & Klimley, 2013).

Combined with other sensors, accelerometers provide a powerful tool to understand the relationships between animal behavior, energetics, and the environment. Many GPS tracking studies infer animal behavior from path geometry, collecting locations at very high intervals to obtain detailed tracks to support inferences about animal behavior based on path trajectories (Grémillet et al., 2004; Mendez et al., 2017; Ryan, Petersen, Peters, & Grémillet, 2004; Wakefield, Phillips, & Matthiopoulos, 2009; Weimerskirch, Le Corre, & Bost, 2008). Pairing GPS and accelerometer sensors could reduce the frequency of required GPS fixes, extending the battery life for longer deployments without sacrificing detailed behavioral data. Satellite and light-based tracking methods record locations with low temporal resolution (geolocators) and at irregular intervals (satellite transmitters), which precludes inference about detailed behavior. If these methods were coupled with accelerometers, then it would be possible to track species over large spatial scales for extended time-periods with high temporal resolution. This type of detailed, long-term tracking of animal movements and behaviors will allow robust inference about animal ecology and how species interact with their environments (Cagnacci et al., 2010; Wakefield et al., 2009).

The ease with which biologists can deploy tracking devices to study the movements of wild animals has exceeded the ability of biologists to categorize, analyze, and interpret the volume of data these efforts have generated. Widespread adoption of accelerometers to measure animal behavior is inhibited by limited validation, which has contributed to a lack of consensus on analysis methods. A host of methods have been proposed for classifying animal behavior from accelerometer data (Appendix S1), including movement thresholds (Brown et al., 2013; Moreau, Siebert, Buerkert, & Schlecht, 2009; Shamoun-Baranes et al., 2012), histogram analysis (Collins et al., 2015), *k*-means (KM) cluster analysis (Angel, Berlincourt, & Arnould, 2016; Sakamoto et al., 2009), *k*-nearest neighbor analysis (Bidder et al., 2014), classification and regression trees (Shamoun-Baranes et al., 2012), neural networks (NN; Nathan et al., 2012; Resheff, Rotics, Harel, Spiegel, & Nathan, 2014), random forests (Bom, Bouten, Piersma, Oosterbeek, & van Gils, 2014; Nathan et al., 2012; Pagano et al., 2017), hidden Markov models (HMM; Leos-Barajas et al., 2016), expectation maximization (EM; Chimienti et al., 2016), and super machine learning (Ladds et al., 2017). At least three custom software applications are available for classifying animal behavior from trained accelerometer data: AcceleRater (Resheff et al., 2014), G-sphere (Wilson et al., 2016), and Ethographer (Sakamoto et al., 2009). Many of these methods use machine-learning techniques that are difficult to interpret because underlying processes are opaque. Numerous accelerometer-derived metrics have been employed as predictors in classification models, often without any critical evaluation of their value in improving classification accuracy. We reviewed 15 similar studies that classified animal behavior from accelerometers, to identify common accelerometer metrics used in classifications (Appendix S1). These studies used between 1 and 147 different variables in their classification models; the median number of parameters included was seven. Using large numbers of predictor variables may make classifications unnecessarily complex, potentially discouraging biologists from adopting this tool, and make methods developed on one data set less generalizable to other studies. Simpler approaches may appear inadequate in comparison to sophisticated analyses, while many complex methods can be difficult for most ecologists to implement.

Identifying an appropriate classification technique is further complicated because most methods are based on small sample sizes, with limited or no validation of classification accuracy. In a sample of 15 studies, only 10 attempted to validate their classifications, only six had sample sizes of more than 10 individuals from the same species, and five studies used data from <5 individuals from some species for analysis (Appendix S1). Many classification methods rely on training data acquired through direct observation of free-living (Nathan et al., 2012), domesticated (Moreau et al., 2009), or captive (Pagano et al., 2017) animals. Training data can be challenging or impossible to collect for wide-ranging species like seabirds, with some species travelling hundreds of kilometers in a single foraging trip. Observations of captive animals are unlikely to represent the full range of animal behavior for species that move over large spatial scales (Pagano et al., 2017). There is a need for robust unsupervised classification methods and for alternative approaches to developing training and validation data sets for species, such as most seabirds, that cannot be observed directly in the wild.

We compared six different methods for classifying behavior using accelerometer data from two seabird species: thick-billed murres (*Uria lomvia*) and black-legged kittiwakes (*Rissa tridactyla*). In this study, we focus on comparing methods for classifying the main behaviors (flying, swimming, on colony, and diving) that comprise a daily activity budget for two seabird species. Daily activity budgets have been widely used in studies of seabird behavior (Ropert-Coudert et al., 2004), energetics (Birt-Friesen, Montevecchi, Cairns, & Macko, 1989), and ecology (Furness & Camphuysen, 1997); identifying robust methods for calculating daily activity budgets from accelerometer data should contribute to wider application of this technology. Accelerometer deployments were paired with GPS data loggers and GPS tracks were used to validate the accuracy of accelerometer-based classifications. High-resolution GPS data are already widely used for behavioral classification in free-living birds, thus, these data provide a good option for validating classifications on a large number of individuals engaging in a full range of natural activities. Our analysis focused on identifying coarse-scale behaviors: resting on colony, flying, swimming, and diving (for murres). Quantifying these behaviors is useful for many seabird studies and these behaviors can be inferred from high-resolution GPS tracks. We compared overall accuracy and behavior-specific accuracy for each species. We also considered the effect of breeding stage (incubation vs. chick rearing) on classification accuracy; although behavior in general should not change between breeding stages, the frequency of different behaviors can change, and factors such as level of activity and posture while at the nest could change, affecting our ability to accurately identify these behaviors. To determine if classification method affects estimates of energy expenditure we also used daily activity budgets from each classification to calculate daily energy expenditure (DEE). Finally, we used variable selection to assess whether or not models using more predictor variables perform better than models with fewer variables and to identify the variables that make the greatest contribution to improvements in classification accuracy for each species.

2 | METHODS

2.1 | Tagging methods

We deployed GPS-accelerometers (Axy-trek; Technosmart, Rome, Italy; 18 g) on 21 incubating and 19 chick-rearing murres breeding at Coats Island, in 2018. Murres were captured using a noose pole and biologgers were attached to the back feathers using TESA tape (TESA 4651, Hamburg, Germany). Murres were released at the capture site and re-captured between 2 and 4 days later to retrieve data loggers. The biologgers were programed to collect GPS locations at 1 min intervals, depth at 0.1 m resolution and 1 Hz intervals, acceleration in three axes at 25 Hz, and temperature at 1 Hz. Note that deployment of similar tags altered dive duration, flight costs, and chick feeding rates (Elliott, Davoren, & Gaston, 2007; Elliott, Vaillant, et al., 2014). As all individuals should be similarly impacted, these tag effects should not affect the results of this study.

We deployed tri-axial accelerometers (Axy-3; Technosmart; 3.2 g), paired with GPS biologgers (CatTraQ; Catnip Technologies, USA; 14 g), on black-legged kittiwakes at Middleton Island, Alaska, USA, in 2013. Data were collected from 17 incubating and 19 chick-rearing kittiwakes. Both biologgers were attached to the back feathers of kittiwakes using Tesa tape (TESA 4651). Kittiwakes were released at the capture site and re-captured between 1 and 3 days later to retrieve data loggers. The biologgers were programed to collect GPS locations at 30 s intervals and tri-axial acceleration at 25 Hz. Deployment of these tags had no impact on reproductive success and survival, but altered flight duration (Chivers, Hatch, & Elliott, 2016). As all individuals should be similarly impacted, these tag effects should not affect the results of this study.

2.2 | Accelerometer-derived metrics

We focused on three types of accelerometer-derived metrics for behavior classifications: wing beat frequency (WBF), pitch, and dynamic acceleration. We chose variables that we thought would be related to the target behaviors based on our prior knowledge of the study species. We calculated WBF by extracting the dominant frequency in the *Z*-axis using a Fast Fourier Transform (FFT) over a 5-s moving window. The FFT was performed using the "fft" function

TABLE 1 Accelerometer-derived metrics calculated prior to behavioral classifications. Only pitch, SD_Z, SD_{ODBA}, WBF, and depth were used in classifications, other statistics shown were calculated to obtain final classification parameters

Statistic	Label	Equation	Description
Static acceleration	S_X, S_Y, S_Z	$\frac{\sum X}{n}, \frac{\sum Y}{n}, \frac{\sum Z}{n}$	Average acceleration in each axis, calculated over a 2-s moving window
Pitch	Pitch	$\tan^{-1}\left(\frac{S_X}{\sqrt{S_Y^2+S_Z^2}}\right) * \frac{180}{\pi}$	Vertical orientation of the body angle
Dynamic acceleration	$D_X, D_Y, D_Z,$	$S_X - X, S_Y - Y, S_Z - Z$	Residual acceleration in each axis, calculated over a 2-s moving window
Overall dynamic body acceleration	ODBA	$\lvert D_X\rvert + \lvert D_Y\rvert + \lvert D_X\rvert$	Dynamic acceleration summed across all three axes
Standard deviation of dynamic acceleration in *Z*-axis	SD_Z	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Dz_i - \frac{\sum Dz}{n}\right)^2}$	Variation in the dynamic acceleration in the *Z*-axis
Standard deviation of overall dynamic body acceleration	SD_{ODBA}	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{ODBA}_i - \frac{\sum \mathrm{ODBA}}{n}\right)^2}$	Variation in the dynamic acceleration in the ODBA
Wing beat frequency	WBF		Dominant frequency in the *Z*-axis, calculated using a 5-s moving window
Depth	Depth		Meters below sea level

in base R. The peak frequency in the *Z*-axis can detect signals that represent motion other than flying (such as walking or sea surface waves), however, for simplicity we refer to this as WBF going forward, as this was the signal we were interested in extracting from the accelerometer. Birds in flapping flight should display characteristic frequencies in vertical motion while travelling.

Pitch measures vertical body angle based on the static acceleration (acceleration averaged over time) of all three axis (Table 1). We expected pitch to change between different behaviors, because the body angle of a bird will change between time on land, swimming, and flight. All pitch values were corrected for differences in device orientation by standardizing acceleration measurements to a pitch of 0° for periods of presumed flight (WBF between 6–9 Hz for murres and 3–6 Hz for kittiwakes) (Elliott, Chivers, et al., 2014), when all birds should have a similar and consistent body orientation (Chimienti et al., 2016; Watanuki et al., 2003).

Dynamic body acceleration integrates the amount of dynamic acceleration (i.e., after removing the static component due to gravity and associated with posture) over a fixed time period, and can be used as an index of movement (Shepard, Wilson, Quintana, et al., 2008). Dynamic body acceleration can be measured along each axis individually, or as a composite of all three axes using overall dynamic body acceleration (ODBA, Table 1). For murres, we used standard deviation of the overall dynamic acceleration, (SD_{ODBA}) as a measure of overall activity level. For kittiwakes, initial data exploration indicated that there was greater relative variability in the *Z*-axis than in the ODBA, therefore, we used standard deviation in the *Z*-axis (SD_Z) to measure activity level.

Table 1 describes the accelerometer metrics calculated from accelerometers; all of these metrics have been used in prior studies classifying animal behavior from accelerometers (Chimienti et al., 2016; Pagano et al., 2017; Shamoun-Baranes et al., 2012). Murre classifications also used depth to identify periods of diving. We calculated pitch and dynamic acceleration using a 2-s moving window (Shepard, Wilson, Halsey, et al., 2008) and WBF using a 5-s window, for both species. Once accelerometer statistics were calculated, we subsampled all data to 1 s intervals to reduce processing time during classification, and because our behaviors of interest occurred at intervals >1 s. All summary statistics are reported as mean ± *SD*.

2.3 | Accelerometer track segmentation

We used a behavior-based track segmentation approach for classification (Bom et al., 2014; Collins et al., 2015). Cliff-nesting murres and kittiwakes must fly to travel between their nest site and foraging areas at sea, therefore, periods of flying should separate colony behavior from swimming behavior. For murres, dives are separated from flights by periods of swimming. We used this prior knowledge of seabird behavior to segment tracks into periods of consistent behavior. We first classified diving (murres) and flying (murres and kittiwakes) from the 1-s sampled data using each method. Any behavior that occurred for <3 s was re-assigned to the previous behavior class and each period of presumed behavior was assigned a unique segment ID. For practical reasons, we imposed a maximum length of 120 s on each segment. This ensured that if a transition between behaviors was missed, the error wold not propagate beyond 120 s. This upper limit also ensured that each type of behavior was represented proportionally in the data. Incubation bouts typically last for many hours, while bouts of flying or diving could last seconds or minutes, so although most of the birds spend a majority of their time at the nest, there would be relatively few bouts of colony behavior relative to other types of behavior. Within each segment, we recalculated movement metrics using mean pitch and mean dynamic acceleration.

2.4 | Accelerometer classification—supervised

We used three supervised classification methods: histogram segregation (HS), random forests (RF), and NN.

2.4.1 | Histogram segregation

We adapted a HS approach from Collins et al. (2015). We used density plots to visualize the distribution of each variable sequentially. Characteristic peaks and valleys in the distribution were used to identify break points for different behaviors. Each behavior was classified using a stepwise approach, once the locations had been assigned to a behavior these locations were not considered for the next variable. We first classified “diving” (murres only) and “flying” using depth and WBF. Accelerometer data were then broken down into segments of continuous behavior and we calculated average pitch and average dynamic acceleration within each segment. Remaining “unknown” segments were classified to “swimming” and “colony” based on peaks in histograms for these two variables. Each track was classified individually.

2.4.2 | Neural network

We used the classifications from the HS method to train the NN models. We did not use the GPS data for training the model because we wanted to test classification approaches that could be applied when GPS data are not available for model training. We randomly chose ten tracks for each species, then, randomly selected 1,000 data points within each behavior class from each of these tracks to make a training dataset. This trained model was used to predict classifications for all tracks within each data set. NN models were run with five hidden nodes using the R Package “nnet,” version 7.3–12 (Venables & Ripley, 2002).

2.4.3 | Random forest (RF)

The random forest (RF) method used the same training data set described above for the NN model. We ran the RF models using the R package “randomForest,” version 4.5-14 (Liaw & Wiener, 2002).

2.5 | Accelerometer classification—unsupervised

We also used three unsupervised classification methods: KM cluster analysis, EM, and HMM. For each method, we ran analysis with between three and six classes and visually examined the classifications to decide on the number of classes that best identified the behaviors of interest. When we identified more than three (kittiwakes) or four (murres) behavior classes, classes were grouped into the behaviors of interest based on expected patterns in behavior.

2.5.1 | k-Means

The KM classification was performed in two steps. For murres, dives were identified manually by classifying all data with depth below −1 m as "diving." A KM classification was performed on WBF to identify two classes, and the class with higher WBF was labelled as "flying." We then segmented all data into bouts of "diving" (murres only), "flying" and "unknown" behavior. Within segments of continuous behavior, we calculated the average pitch and dynamic acceleration. A second KM classification was performed on the remaining "unknown" segments with average pitch and dynamic acceleration as input variables. We used the natural logarithm of dynamic acceleration, and both variables were scaled to their range prior to analysis. The KM classification was performed on all tracks at once. Analysis was run using the "kmeans" function in base R.

2.5.2 | Expectation maximization

The EM classification was performed in two steps. For murres, dives were identified manually by classifying all data with depths below −1 m as "diving." An EM classification was performed on WBF to identify two classes; the class with higher WBF was labelled as "flying." We then segmented all data into bouts of "diving" (murres only), "flying" and "unknown" behavior. Within segments of continuous behavior, we calculated the average pitch and dynamic acceleration. A second EM classification was performed on the remaining "unknown" segments, with average pitch and dynamic acceleration as input variables. We used the natural logarithm of dynamic acceleration, and both variables were scaled to their range prior to analysis. EM classification was performed on all tracks for each species at once EM analysis was conducted using the R package "Rmixmod" package, version 2.1.1 (Langrognet, Lebret, Poli, & Iovleff, 2016). We considered Gaussian models with free proportions; BIC was used to identify the best model.

2.5.3 | Hidden Markov models

Hidden Markov models require data that are sampled at equal intervals, for this reason, we did not use the track segmentation approach described above. Instead, average accelerometer values for WBF, pitch, dynamic acceleration and depth were taken for 5-s intervals (murres) and 10-s intervals (kittiwakes). A shorter interval was used for murres to preserve short inter-dive bouts. We used the R package "momentuHMM" (McClintock & Michelot, 2018) to fit HMMs. For murres, depth data were converted to a binary variable, where data with depth <−1 m received a value of 1 and depths >−1 received a value of 0, this was modelled using a Bernoulli distribution. A full description of the distributions and starting values used for each behavior and variable is provided in Tables 2 and 3. We fixed transition probabilities between colony-swimming, swimming-colony, colony-diving, diving-colony, diving-flying, and flying-diving to zero. The most likely behavioral states were obtained from the model using the Viterbi algotrithm (McClintock & Michelot, 2018).

2.6 | GPS classification

2.6.1 | Thick-billed Murre

We used GPS and depth sensor data to validate murre behavior classifications. Locations requiring a calculated ground speed >30 m/s were excluded from analysis (0.1% of all GPS locations), because these were potential GPS errors. If depth was below −1 m, data were labelled as diving. Remaining locations with a calculated ground speed >2 m/s were classified as flying. At relatively high sampling rates (i.e., <100 s), like those used in this study, the calculated ground speed and instantaneous speeds are expected to be highly correlated (Elliott, Chivers, et al., 2014). Locations within 250 m of the nest were classified as colony and all remaining locations were classified as swimming. Following this initial classification, each bout of continuous behavior was assigned a unique identifier. Data were examined for obvious classification errors based on temperature,

TABLE 2 Starting values for the state-dependent probability distribution parameters for variables used in the hidden Markov model to classify behavior of thick-billed murres

Variable	Family	Link	Parameter	Colony	Diving	Flying	Swimming
Pitch	Normal	Identity	Mean	30	−5	0	−5
		Log	*SD*	20	50	5	10
SD_{ODBA}	Exponential	Log	Rate	25	5	2.5	5
WBF	Log normal	Identity	Location	0.5	2	9	2
		Log	Scale	0.5	0.5	0.2	0.5
		Logit	Zero-mass	0.9	0.9	0.1	0.9
Depth	Bernoulli	Logit	Probability	1×10^{-12}	$1 - (1 \times 10^{-12})$	1×10^{-12}	1×10^{-12}

TABLE 3 Starting values for the state-dependent probability distribution parameters for variables used in the hidden Markov model to classify behavior of black-legged kittiwakes

Variable	Family	Link	Parameter	Colony 1	Colony 2	Flying	Swimming
Pitch	Normal	Identity	Mean	35	10	0	5
		Log	*SD*	10	10	5	5
SD_Z	Log normal	Identity	Location	0.05	0.05	0.6	0.15
		Log	Scale	0.5	0.5	0.5	0.5
		Logit	Zero-mass	0.9	0.9	0.1	0.1
WBF	Log normal	Identity	Location	0.5	2	9	2
		Log	Scale	0.5	0.5	0.2	0.5
		Logit	Zero-mass	0.9	0.9	0.1	0.9

duration of behavior, and behavioral context (prior and subsequent behaviors). Swimming bouts within 3 km of the colony with a high average temperature (>10°C) were examined as potential colony locations and colony bouts with low average temperature were examined as potential swimming locations. Only 0.6% of GPS locations were manually reclassified.

Because the GPS data were collected at a lower temporal resolution (60 s for murres and 30 s for kittiwakes) than the accelerometer analysis (1 s), the GPS classification would be slower to respond to a change in behavior. For example, a murre that transitions from flying to swimming halfway between two GPS fixes would be classified as still flying during the next location, however the accelerometer could pick up this change in behavior at the time it occurred. To deal with this difference in sampling rate, we identified periods when the GPS indicated a transition from one behavior to another. All data points within 60 s of a GPS transition between colony, flying, or swimming were labelled as transitions and excluded from further analysis. Transitions between diving and swimming were not excluded, because the pressure sensor collected depth data at 1 s intervals. In total, 11.0% of GPS locations were excluded for murres because they were identified as periods of transition between behaviors.

2.6.2 | Black-legged Kittiwake

GPS data were used to validate kittiwake behavior classifications. Locations requiring a ground speed >20 m/s or more than 10-min between fixes were excluded from analysis (0.4% of locations), because these were potential GPS errors. Locations with a calculated ground speed >3 m/s were classified as flying. Locations on Middleton Island were classified as colony, and all remaining locations were classified as swimming. Kittiwakes can spend significant time on tidal flats and sand bars around Middleton Island (K. Elliott, personal observation); in these locations, birds could be swimming or loafing depending on tide heights and these behaviors could not be differentiated based on the GPS data alone. Therefore, we excluded all locations within 500 m of the island from the analysis. This reduced the total GPS data set by 11.1%; this step was important to minimize uncertainty and potential for errors in our validation data. Similar to murres, all locations within 30 s of a change in behavior were labelled as transitions (13.5%) and excluded from the analysis.

2.7 | Classification accuracy

We subsampled the accelerometer data to 1 min (murres) and 30 s (kittiwakes) to match the resolution of the GPS data and used a confusion matrix to calculate the overall accuracy and the balanced accuracy for each behavior. Confusion matrices and measures of accuracy were calculated using the R package *carat* (Kuhn, 2016). We used mixed-effects models, with bird identity as a random effect, to test for differences in the classification accuracy among methods and between breeding stages. Accuracy data were logit transformed prior to analysis. We used the R package *nlme* (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2018) to run the models and the *lsmeans* package (Lenth, 2016) to calculate parameter estimates, 95% confidence intervals (CI) and for pairwise comparisons.

2.8 | Daily energy budget

We used an estimate of DEE to look at the overall variation among classification methods. DEE (in kJ/d) for murres was calculated following Elliott et al. (2013) as:

$$\mathrm{DEE} = 32.0 * t_c + 532.8 * t_f + 100.8 * t_s + 97.2 * t_d.$$

Daily energy expenditure for kittiwakes was calculated following Jodice et al (2003), using activity specific metabolic rates for nest attendance, commuting flight, and surface feeding to develop the equation:

$$\mathrm{DEE} = 21.0 * t_c + 99.9 * t_f + 25.8 * t_s,$$

where *t* is time in hours and the subscripts are c = colony, f = flying, s = swimming, and d = diving. We converted metabolic rates from CO_2 production rates (ml CO^2 $g^{-1-}hr^{-1}$) to kJ using an energetic equivalent of 27.33 kJ L CO_2 assuming average kittiwake mass of 416 g (Jodice et al., 2003; Speakman, 1997). We used mixed effects models, with bird ID as random effects, to test for differences in DEE estimates among methods.

2.9 | Variable selection

We chose 42 accelerometer statistics used in previous studies (Appendix S1) to consider in our variable selection analysis; these included raw acceleration values, static acceleration, dynamic acceleration, minimum, maximum, range, skew, and kurtosis for each axis. We also calculated the trend, as the slope coefficient from a linear regression, and autocorrelation, as the value of the first order autocorrelation function. Each of these statistics was calculated over a 2-s moving window. Finally, we included the dominant frequency for each axis calculated over a 5-s moving window.

We used random forests models to identify which variables contributed the most to classification accuracy and how much adding additional variables improved accuracy. To simulate a realistic training data set, acquired through paired GPS-accelerometer deployments, we trained and tested data from the classified GPS tracks using a random subset of 10 individual tracks for each species. From these tracks, we sub-sampled 1,000 locations from each behavior class to ensure each behavior was adequately represented in the training data. We used forward selection to identify which accelerometer variables provided the greatest improvement in classification accuracy for models with between 1 and 10 variables. To reduce overall computation time, variable importance from a global model with all variables and all training data were used to identify the 20 most influential variables to include in the variable selection analysis. At each step, we ran 100 simulations with randomly selected training data sets and selected the variable with highest median accuracy over all simulations. We compared model accuracy among the best models with 1–10 variables and a global model with all 42 variables. Confidence intervals for model accuracy are based on the 2.5th and 97.5th percentile of all simulations.

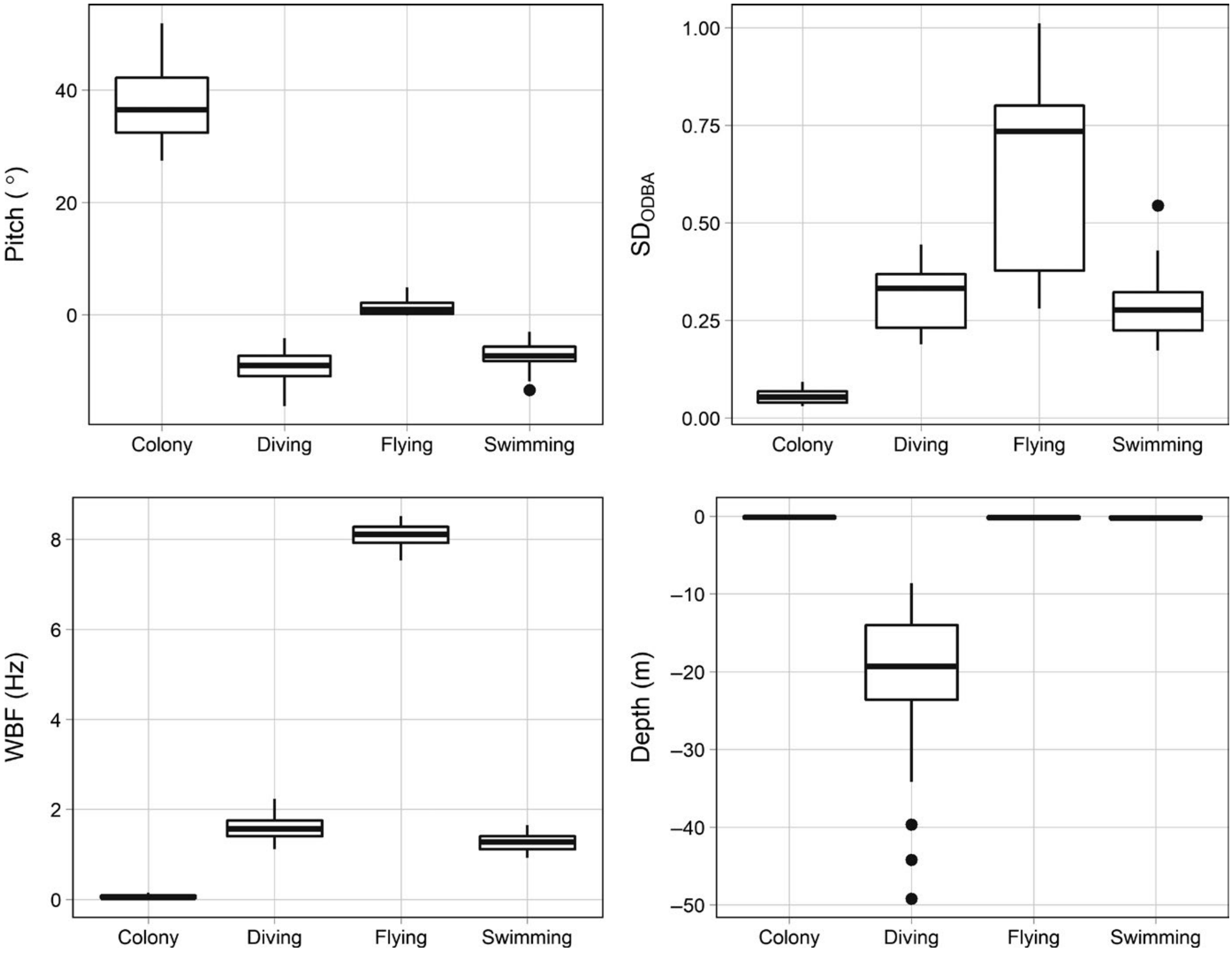

FIGURE 1 Boxplots showing the distribution of average values of predictor variables for each thick-billed murre behavior

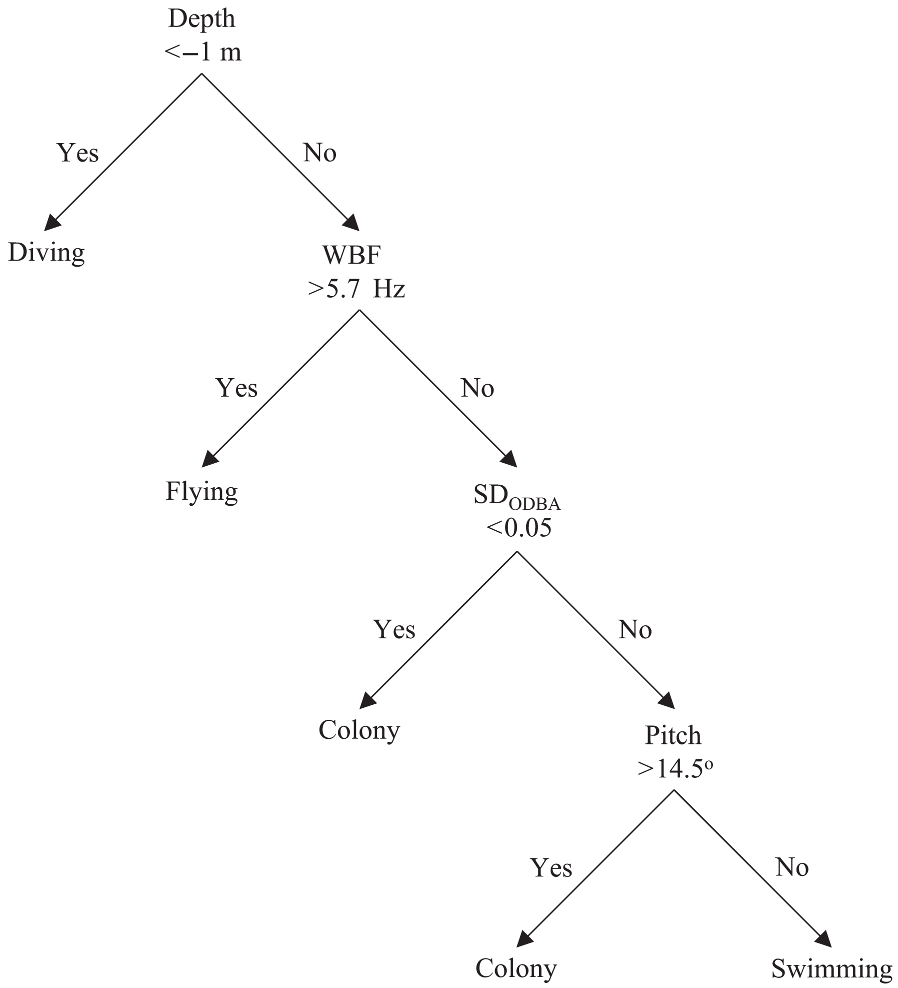

FIGURE 2 Diagram showing the average break points and classification hierarchy used in the histogram segregation method for thick-billed murres

3 | RESULTS

3.1 | Classification summary

3.1.1 | Murres

Colony segments were characterized by high pitch (37.6 ± 6.1°; Figure 1) and low SD_{ODBA} (0.05 ± 0.02 g). Swimming segments were characterized by low pitch (−7.4 ± 2.5°) and high SD_{ODBA} (0.28 ± 0.08 g). Flying segments had high WBF (8.1 ± 0.25 Hz). Diving segments were characterized by depths below −1 m (−20.5 ± 9.0 m). Figure 2 shows the hierarchical process and average breakpoints used for assigning behaviors with the HS method. We used five total classes in the KM classification for murres: two colony, one diving, one flying, and one swimming class. For the EM and HMM classes only four classes were necessary to obtain a clear separation of all four behaviors, based on visual examination of the classifications.

3.1.2 | Kittiwakes

Colony segments were characterized by high pitch (29.9 ± 11.7°; Figure 3) and low SD_Z (0.04 ± 0.02 g). Swimming was characterized by low pitch (5.7 ± 2.9°) and high SD_Z (0.18 ± 0.04 g). Flying segments had high WBF (4.16 ± 0.16 Hz). The HS method began by classifying flight with WBF, then colony with SD_Z, and finally swimming with pitch (Figure 4). We used four total classes in the KM, EM, and HMM classifications for kittiwakes: two colony classes, one flying class, and one swimming class.

3.2 | Classification accuracy

3.2.1 | Murres

Mean classification accuracy for each method was >98.3% and accuracy for each individual track was above 92.7% for all methods (Figure 5). There was no statistical support for a difference in accuracy among classification methods ($F_{5,190}$ = 1.28, p = 0.28). Averaging across breeding status, accuracy was highest using the HS (98.5%; CI = 98.1–98.7) method and lowest for the HMM (98.3%; CI = 97.9–98.6) method, but this difference was not statistically significant (t_{190} = 2.162, p = 0.26). Accuracy for all methods was higher for murres with chicks (98.4%, CI = 97.9–98.8) than for murres with eggs (98.2%, CI = 97.7–98.6); however, there was no evidence that

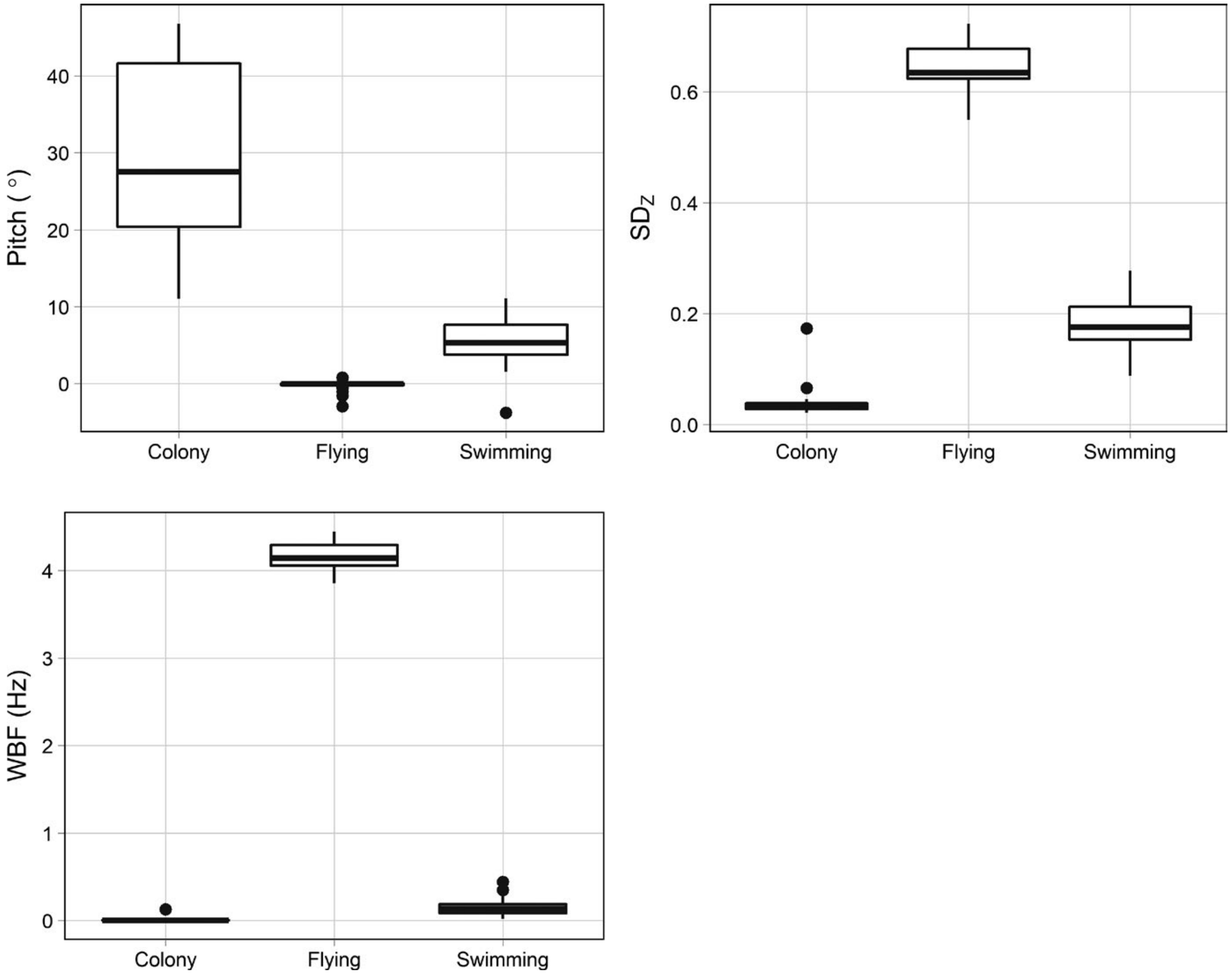

FIGURE 3 Boxplots showing the distribution of average values of predictor variables for each black-legged kittiwakes behavior

accuracy varied with breeding status ($F_{1,38}$ = 0.46, p = 0.50) or for an interaction between method and breeding status ($F_{5,190}$ = 0.75, p = 0.58).

There was a significant interaction between method and behavior ($F_{15,894}$ = 23.6, $p < 0.001$; Figure 6), indicating that some methods were more accurate at classifying certain behaviors than other methods. Average classification accuracy for colony across all methods was 99.1% (CI = 99.5–99.7); there were no significant differences in classification accuracy among methods for colony (all $p > 0.35$). Average classification accuracy for swimming across all methods was 98.7% (CI = 98.4–98.9); there were no significant differences in classification accuracy among methods for swimming (all $p > 0.06$). The HMM method was most accurate for classifying flying (97.9%, CI = 97.4–98.3); this was significantly higher than all other methods (NN: 95.3%, CI = 94.3–96.2; t_{894} = 6.88, $p < 0.001$; HS: 95.3%, CI = 94.3–96.2; t_{894} = 6.93, $p < 0.001$; RF: 95.3%, CI = 94.3–96.2; t_{894} = 6.97, $p < 0.001$; EM :94.4%, CI = 93.1–95.4; t_{894} = −8.58, $p < 0.001$; KM: 94.3%, CI = 93.1–95.4; t_{894} = 8.69, $p < 0.001$). For diving, classification accuracy was highest for the HS (99.9%; CI = 99.8–1.00), EM (99.9; CI = 99.8–1.00), and KM (99.9%; CI = 99.8–1.00) methods, and lowest for the HMM method (98.2%; CI = 97.8–98.6). High classification accuracy for diving is expected, because dives in the GPS data and accelerometer data were both classified using the depth sensor. There was a significant interaction between behavior and stage ($F_{3,894}$ = 15.9, $p < 0.001$). Flying was classified more accurately during chick rearing (96.7%, CI = 95.9–97.4) than during incubation (94.2%, CI = 92.9–95.2; t_{38} = 3.92, $p < 0.001$) and there was some evidence that swimming was classified more accurately during chick rearing than incubation (t_{38} = −1.91, p = 0.06).

3.2.2 | Kittiwakes

There was strong evidence for a difference in classification accuracy among methods ($F_{5,170}$ = 6.21; $p < 0.001$) and between breeding stages ($F_{1,34}$ = 9.41; p = 0.004), there was no support for an interaction between method and breeding stage ($F_{5,170}$ = 0.41; p = 0.84; Figure 5). Averaging across all methods, accuracy during the chick stage was 93.7% (CI = 92.0–95.0) while accuracy was 89.5% (CI = 86.7–91.7) during the egg stage. For birds with eggs, there were no significant differences among the HMM (90.3%, CI = 87.7–92.5), HS (90.0%, CI = 87.3–92.3), EM (89.6%, CI = 86.7–91.9), RF (89.4%, CI = 86.5–91.8), and NN (89.1%, CI = 86.2–91.5) methods. The KM (88.2%, CI = 85.1–90.8) method was significantly less accurate than the HMM (t_{170} = −3.87, p = 0.002) and HS (t_{170} = −3.25, p = 0.02) methods. The absolute difference in accuracy between the most accurate method, HMM, and the least accurate method, KM, was only 2.1%. During the chick rearing stage, there were no differences in classification accuracy among the HMM (94.2%, CI = 92.6–95.4), RF (93.7%, CI = 92.0–95.1), NN (93.7%, CI = 92.0–95.1), HS (93.7%, CI = 92.0–95.1), and EM (93.6%, CI = 91.8–95.0) methods. Classification accuracy for the KM method (93.0%, CI = 91.0–94.5) was significantly lower than the HMM (t_{170} = −3.75, p = 0.003) method. The absolute difference in accuracy between the best and worst classification methods was only 1.2%.

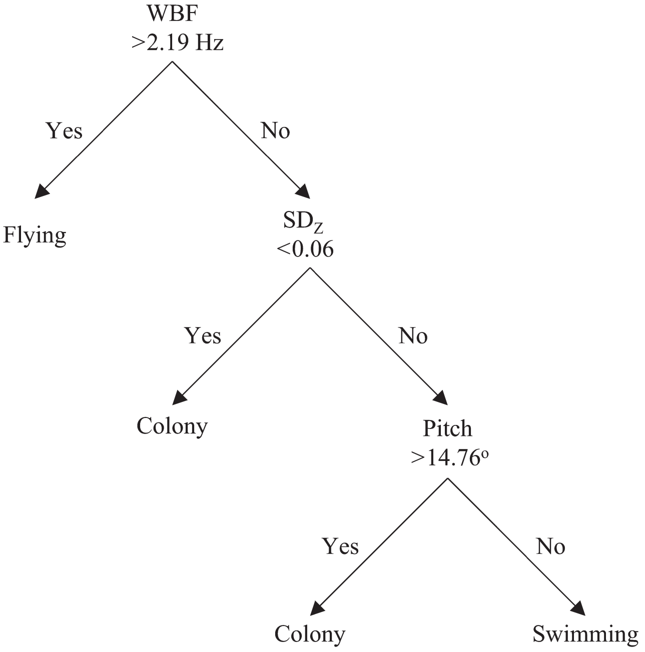

FIGURE 4 Diagram showing the average break points and classification hierarchy used in the histogram segregation method for black-legged kittiwakes

There was no interaction between method and behavior ($F_{10,593}$ = 0.66; p = 0.77), indicating that all methods classified different behaviors with similar accuracy. There was a significant interaction between behavior-specific accuracy and breeding stage ($F_{2,593}$ = 163.0; p < 0.001; Figure 6). Colony behavior was identified more accurately during the chick stage (97.6%, CI = 97.1–98.1) than during the egg stage (90.0%, CI = 87.8–91.8; t_{34} = −10.3; p < 0.001). There was no difference in classification accuracy for swimming across stages (Eggs: 92.2%, CI = 90.5–93.7; Chicks: 93.1%, CI = 91.7–94.4; t_{34} = −1.632, p = 0.11). There was also no difference in accuracy of flight classification between stages (Eggs: 88.5%, CI = 83.0–88.3; Chicks: 88.5%, CI = 86.3–90.5; t_{34} = −0.92, p = 0.37).

3.3 | Daily energy budget

3.3.1 | Thick-billed murres

There was a significant difference in estimates of DEE among methods ($F_{5,190}$ = 40.3, p < 0.001) and suggestive evidence of an interaction between method and breeding status ($F_{5,190}$ = 2.19, p = 0.06). For murres with eggs, mean DEE calculated with the RF classification (2,112 kJ/day, CI = 1,908–2,315) was lower than DEE calculated with all other methods (EM 2,242 kJ/day, CI = 2,038–2,446, t_{190} = −8.76, p < 0.001; HS: 2,242 kJ/day, CI = 2,038–2,446, t_{190} = −8.76, p < 0.001; KM: 2,242 kJ/day, CI = 2,038–2,446,

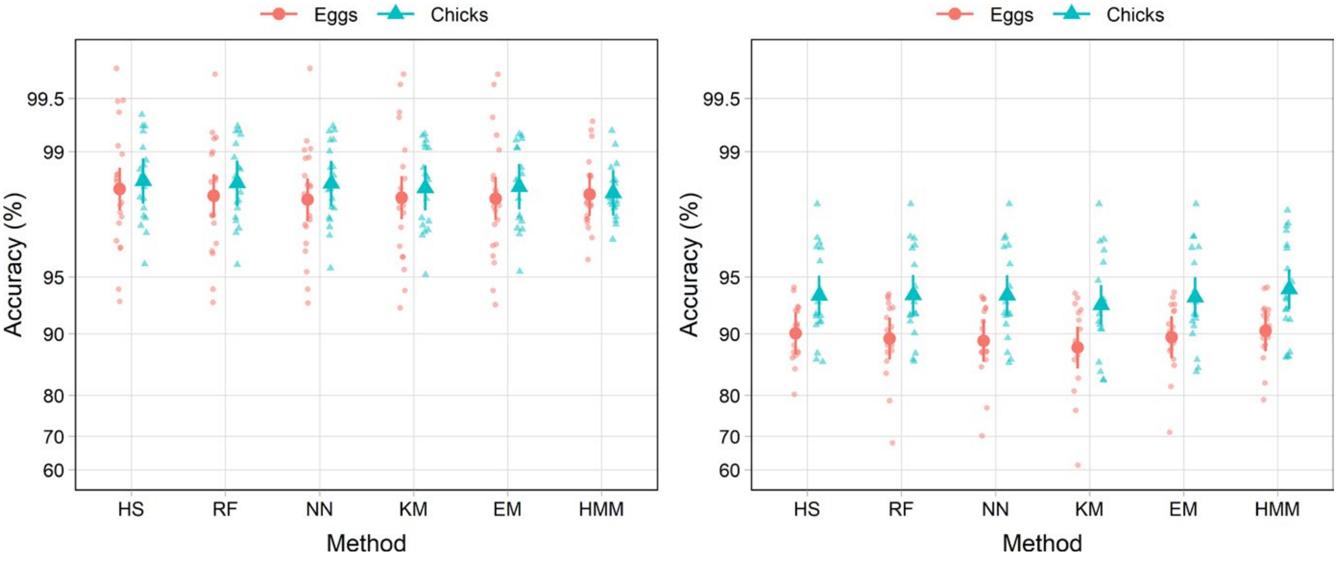

FIGURE 5 Average accuracy of classification methods for thick-billed murre (left) and black-legged kittwakes (right). Large symbols show group means and error bars are 95% confidence intervals, small symbols are data from each individual. Data are displayed on a logit scale

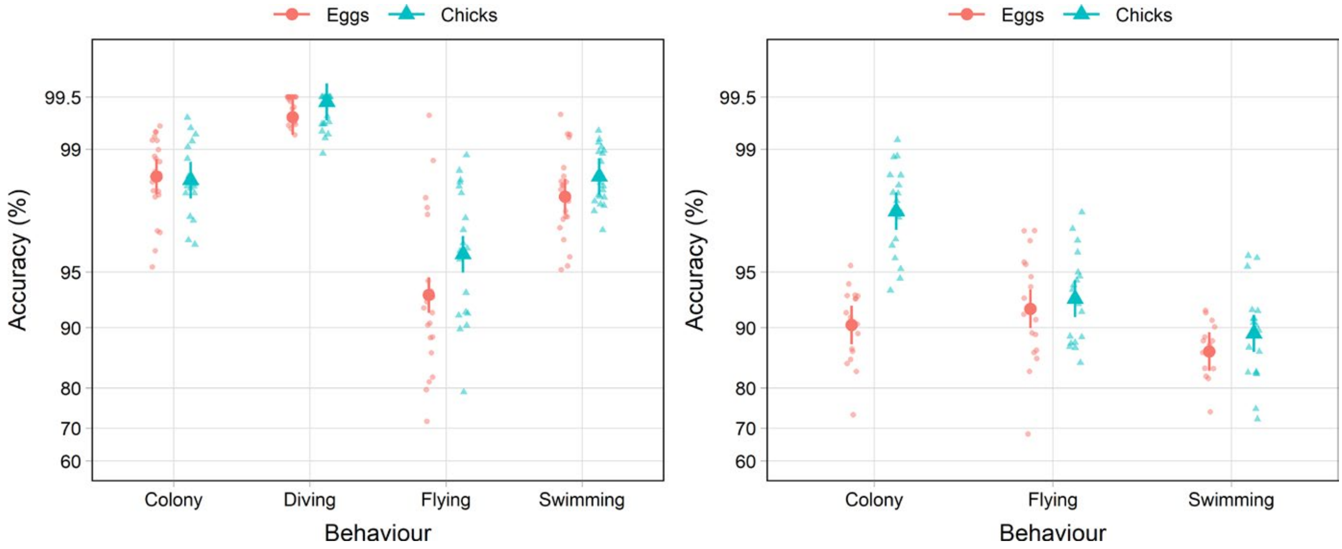

FIGURE 6 Average accuracy for thick-billed murre (left) and black-legged kittwakes (right) behaviors; only results from the histogram segregation (HS) method are shown. Large symbols show group means and error bars are 95% confidence intervals, small symbols are data from each individual. Data are displayed on a logit scale

t_{190} = −8.77, p < 0.001; HMM: 2,265 kJ/day, CI = 2,061–2,468, t_{190} = −10.3, p < 0.001; NN: 2,272 kJ/day, CI = 2,069–2,476, t_{190} = −10.8, p < 0.001). During incubation, the difference between average DEE estimate using the RF method and the NN method, was only 161 kJ or 7.1% of mean DEE. During chick rearing, mean DEE calculated using the RF (2,375 kJ/day, CI = 2,161–2,589) classification was significantly lower than all other methods (KM 2,454 kJ/day, CI = 2,240–2,669, t_{190} = −5.06, p < 0.001; EM: 2,455 kJ/day, CI = 2,240–2,669, t_{190} = −5.06, p < 0.001; HS: 2,455 kJ/day, CI = 2,241–2,669, t_{190} = −5.11, p < 0.001; HMM: 2,471 kJ/day, CI = 2,257–2,685, t_{190} = −6.11, p < 0.001; NN: 2,475 kJ/day, CI = 2,260–2,689, t_{190} = −6.36, p < 0.001). The difference between average DEE estimate during chick rearing using the RF method and the NN method, was only 99 kJ or 4.0% of mean DEE.

3.3.2 | Kittiwakes

Breeding status had a significant effect on DEE ($F_{1,37}$ = 23.5, p < 0.001); kittiwakes with chicks (1,222 kJ/day, CI = 1,116–1,327) had significantly higher DEE than kittiwakes with eggs (869 kJ/day, CI = 767–972). Classification method had a significant effect on estimates of DEE ($F_{5,185}$ = 74.8, p < 0.001). During incubation, the RF method had significantly lower estimates of DEE (842 kJ/day, CI = 739–944) than all other methods (NN: 874 kJ/day, CI = 771–977, t_{185} = −9.09, p < 0.001; KM: 875 kJ/day, CI = 772–978, t_{185} = −9.31, p < 0.001; HS: 875 kJ/day, CI = 772–978, t_{185} = −9.32, p < 0.001; HMM: 875 kJ/day, CI = 772–978, t_{185} = −9.35, p < 0.001; EM: 876 kJ/day, CI = 773–977, t_{185} = −9.63, p < 0.001). However, the difference between average DEE estimates during incubation using the RF method and the EM method, which had the highest average DEE estimates, was only 34 kJ or 3.9%. For kittiwakes with chicks, the RF method (1,185 kJ/day, CI = 1,080–1,291) also had significantly lower estimates of DEE than all other methods (HMM: 1,229 kJ/day, CI = 1,123–1,334, t_{185} = −9.09, p < 0.001; HS: 1,229 kJ/day, CI = 1,123–1,334, t_{185} = −9.31, p < 0.001; KM: 1,229 kJ/day, CI = 1,223–1,334, t_{185} = −9.32, p < 0.001; EM: 1,229 kJ/day, CI = 1,123–1,334, t_{185} = −9.35, p < 0.001; NN: 1,229 kJ/day, CI = 1,224–1,335, t_{185} = −9.63, p < 0.001). During chick rearing, the difference between average DEE estimate using the RF method and the NN method, was only 44 kJ or 3.6% of mean DEE.

3.4 | Variable selection

3.4.1 | Thick-billed murres

Classification accuracy increased from the best possible model using a single variable, 81.0% (CI = 78.7–82.3) to 98.7% (CI = 98.2–98.9) accuracy for the best model using three variables. Adding more than three variables to the model did not increase model accuracy (Figure 7). Variable selection identified WBF, depth, and static_X, as the three variables with the greatest influence on classification accuracy. A global model using all 43 candidate variables had 98.8% (CI = 98.2–99.1) classification accuracy, which overlaps the accuracy achieved with the three-variable model. Following the same procedure using our original variables, WBF, pitch, depth, and sdODBA, gave comparable accuracy at 98.5% (97.7–98.9). Pitch, one of our a priori variables, was the fifth variable after static_X and skew_Z. Static_X and pitch had a correlation coefficient of 0.96 (CI = 0.964–0.965; p < 0.001); therefore, these two variables may be largely interchangeable. Our chosen measure of dynamic acceleration, SD_{ODBA}, did not rank among the twenty most important variables, indicating that including this metric in our original models may not have contributed to classification accuracy.

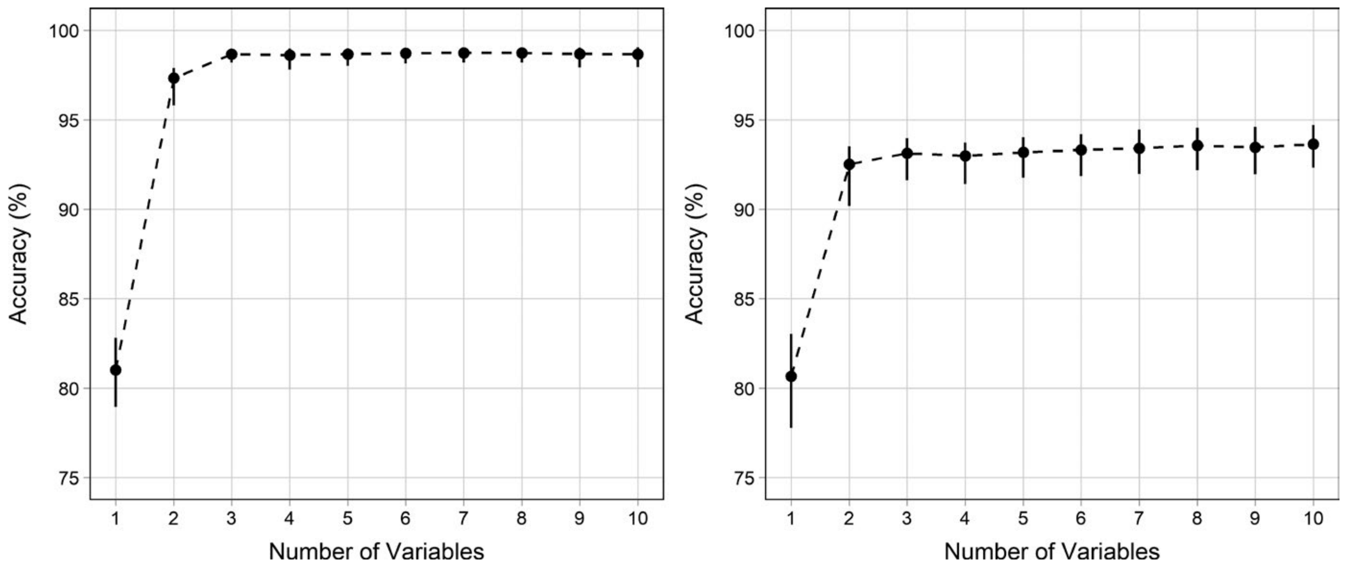

FIGURE 7 Change in thick-billed murre (left) and black-legged kittiwake (right) behavior classification accuracy with additional variables included in random forest models using a forward selection procedure. Black points are medians and error bars are 95% confidence intervals

3.4.2 | Kittiwakes

Classification accuracy increased from the best possible random forest model using a single variable, 80.7% (CI = 77.8–83.0), to the best model using two variables, 92.5% (CI = 90.2–93.5). Additional variables did not substantially increase model accuracy (Figure 7). A global model using all 42 candidate variables had 93.4% (CI = 91.7–94.6) classification accuracy, which overlaps the accuracy achieved with the best two-variable model. Forward selection identified autocorrelation in the *Z*-axis (ACF_z) and WBF as the best predictors in a two-variable model. ACF_Z had low values during colony segments (0.1 ± 0.1), intermediate values during flying segments (0.5 ± 0.04), and high values during swimming segments (0.7 ± 0.12). As with the initial classification methods, WBF was high during periods of flight and low during periods of swimming or periods on the colony.

Our original model using WBF, pitch and SD_Z had comparable accuracy, 92.5% (CI = 90.4%–93.6%), to the top two variable model identified through variable selection. ACF_Z appeared to measure differences in activity in kittiwake behavior that were not apparent in pitch or SD_Z. For both pitch and SD_Z, average values of pitch and SD_Z for colony were more similar to swimming than flying, while average values of ACF_Z for colony and swimming were more distinct than from average values for flying. Since our original model had lower accuracy for swimming and colony behavior, at least during incubation, ACF_Z may provide better classification for these behaviors.

4 | DISCUSSION

We found high classification accuracy using a small number of accelerometer-derived metrics to identify coarse-scale animal behavior. Accuracy was robust to choice of classification method. Although there were statistically significant differences in classification accuracy for the methods tested, average accuracy of all methods was high (98% murres, 91% kittiwakes). There were no differences in mean accuracy among methods for murres and relatively small differences in mean accuracy among methods for kittiwakes. Choice of classification method appears to have little impact on classification results. Any of the methods described here should provide a robust classification of the principal behavior types for murres and kittiwakes. We expect these results to be largely transferable to other species in the same families, and potentially more broadly applicable to other waterbirds that use flapping flight.

We were able to achieve highly accurate and consistent results across all methods using a small set of predictor variables. For both species, including more than two or three predictor variables gave no significant improvement in classification accuracy. Many other studies, particularly those using machine learning methods, include large numbers of predictor variables (Ladds et al., 2017; Nathan et al., 2012). We found that limiting the number of variables greatly reduced analysis time, because files are smaller and models are simpler. Resulting classifications are easier to interpret, especially for unsupervised classifications, because they are based on fewer predictors with an a priori relationship to behavior.

More importantly, we have shown that similar variables–pitch, dynamic acceleration, and WBF–can be used to classify the behavior of two different seabird species. The predictor variables we selected are likely to be useful in classifying coarse-scale behaviors for a wide range of species, because changes in pitch, dynamic acceleration, and periodicity are fundamental components of all activity (Shepard, Wilson, Quintana, et al., 2008). Even in non-flying species, locomotion (walking, running, and swimming) should have a distinct signature in the frequency domain which would help identify this type of behavior (Shepard, Wilson, Quintana, et al., 2008). Measures of pitch, dynamic acceleration, and frequency should be a good starting point in any behavioral classification. However, our variable selection identified another variable, ACF_Z for kittiwakes, which performed slightly better in classifying behavior for this species, the difference in average accuracy in using this variable was minimal. In the absence of training data to conduct similar variable selection, the types of accelerometer statistics we selected a priori for our models are likely to be effective in classifying basic behavior for a range of species.

That classification accuracy was consistently high is perhaps not a surprising result. Many studies have found higher accuracy when only a small number of general behaviors is considered (Hammond, Springthorpe, Walsh, & Berg-Kirkpatrick, 2016; Ladds et al., 2017; Shamoun-Baranes et al., 2012). Indeed, the behaviors we considered are readily identifiable in an accelerometer trace using the human eye. The challenge for researchers is developing methods that can automatically, and reliably, label these behaviors. This study is notable because we have demonstrated that these behaviors are easily identifiable using large data set from two different, wide-ranging seabird species, which cannot be easily observed in the wild.

Our classification of murre behavior benefitted from incorporating data from a pressure sensor to measure depth and identify dives. However, the behavior specific accuracy for the other three behaviors (colony, flying, and diving) were all >94%, so even if diving was excluded the overall classification accuracy for murres would have been high using our methods. Pressure sensors add little to the weight and size of an accelerometer, so for most diving species there is no reason not record pressure data along with acceleration. For very small diving species, further development of methods to classify dives and estimate depth using only accelerometer data are needed.

Classification accuracy is not the only factor that should influence choice of classification method. Depending on the research questions being addressed, certain methods may be more appropriate. Hidden Markov models offer advantages, beyond high classification accuracy, that are not achieved with the other methods considered here. Specifically, HMMs account for the serial dependence in an acceleration time series (Leos-Barajas et al., 2016). In this study, we could directly model the expected transitions between our three or four behavioral states by setting priors on the transition probabilities. Indeed, for the other classification methods we used a track segmentation approach to improve our ability to detect broad scale behaviors. Our segmentation approach would not work for species that do not have to transition through one behavior (e.g., flight) to begin another behavior. HMMs can also be used to jointly model how external factors influence behavior (Leos-Barajas et al., 2016). Using other methods, this must be done in as a two-step process, first classifying behavior and then testing for relationships with external factors. However, the HMMs are arguably the least accessible method we considered; they require sophisticated statistical understanding to implement, and success in behavioral classification depends on carefully specified priors. For applications where behavioral classification is likely to be high, and data will ultimately be summarized at large timescales (e.g., hours, days, or longer), the advantages of using HMMs may not outweigh the costs of implementing this method.

Our methods worked across two different species and breeding stages (incubation vs. chick-rearing). Nonetheless, classifications were more accurate with murres than kittiwakes across all methods. Murres have high wing loading and high wing beat frequencies (Elliott et al., 2013; Pennycuick, 1987). As a result, murres only use flapping flight, which is easily defined from accelerometer profiles. Kittiwakes have much lower wing loading and lower wing beat frequencies (Jodice et al., 2006; Pennycuick, 1987). Murres make rapid, directed flights with few landings on the water, which helps to distinguish flight from swimming in GPS tracks. The more agile kittiwakes change direction and make short, frequent landings while visually searching for prey, which would create more overlap in ground speeds measured by GPS. Simultaneous deployments of GPS-accelerometers with salinity loggers or a magnetometer could help improve validation of kittiwake behavior classifications and identify accelerometer measures characteristic of gliding flight.

In principle, there should be no difference in the behaviors we classified between incubation and chick rearing, because all of these behaviors occur in all stages of the annual cycle. However, we did find it was more difficult to classify swimming and colony behavior accurately for incubating kittiwakes than for chick-rearing kittiwakes. For both species, swimming was primarily differentiated from colony using differences in dynamic acceleration and pitch. Kittiwakes build a nest structure to hold their eggs and can be quite active in shifting positions and turning eggs within their nest cup. This activity at the nest and changes in pitch during incubation may have made it more difficult to differentiate incubation from swimming consistently. Additionally, during incubation kittiwakes may spend more time resting on the water, which would have relatively low dynamic acceleration compared to active foraging on the water, making it more difficult to discern from time spent at the nest. Variable selection analysis found that ACFz was a stronger predictor of behavior for kittiwakes than either pitch or SD_Z. ACF_Z showed strong differentiation between swimming and colony, making it potentially a more useful variable in classifications for kittiwakes.

For any behavioral classification, the position of the data logger on the animal could influence the utility of certain acceleration measures. For example, a logger mounted on the tail or legs would have a different pitch signature than a logger mounted on the back or stomach, and may show different patterns of dynamic acceleration from the main body. Additionally, variation in how loggers are attached to individual animals can influence the ability to identify different behaviors between tracks. Indeed, in our data the differences in classification accuracy among individuals was significantly larger than the differences in classification accuracy among methods. Therefore, there should be careful consideration of logger position, and consistency in logger attachment, during study design, implementation, and data analysis.

By using a training data set for the RF and NN methods that only included a sub-sample of individuals, we demonstrated that data from a small number of individuals was transferable to a larger sample of individuals. Acquiring training data for wide-ranging species like seabirds is an impediment to using supervised classification methods for labelling behaviors. We have demonstrated that a simple supervised classification method can be used to build a training data set for basic behaviors in seabirds. The NN and RF approaches have the advantage that classifications can be fully automated without any user input once a training data set has been developed. The use of machine learning techniques for classification of wide ranging species can be limited by the challenges of developing a training data set. With large data sets, a training data set could be developed based on a subsample of data using any of the other four methods described here, and a model based on this training data could be used to classify remaining data.

Wing beat frequency was an important variable in our classifications. Estimating WBF from accelerometer data requires a sampling frequency that is at least two times higher than the expected WBF (or equivalent movement pattern) of the focal species ("the Nyquist frequency"). WBF also has many ecological applications, such as estimating changes in mass after a foraging bout (Sato, Daunt, Watanuki, Takahashi, & Wanless, 2008) and measuring changes in flight costs associated with environmental conditions (Elliott et al., 2013). Flapping flight is one of the most energetically expensive behaviors for seabirds, so accurately quantifying this behavior is important for energetic estimates. We recommend accelerometer studies on seabirds use a sampling frequency that will allow estimation of WBF, which is consistent with other authors recommendations for sampling frequencies to adequately sample dynamic body acceleration (Gómez Laich, Wilson, Gleiss, Shepard, & Quintana, 2011). For behavioral classifications, we cannot perceive any strong rationale for sampling at frequencies higher than 2–3 times the expected WBF of a focal species.

Coarse-scale behavior identification, like the approaches demonstrated here, could be a first step in a hierarchical process of identifying fine-scale behaviors (Leos-Barajas et al., 2017, 2016). Several studies have been successful in distinguishing general behaviors, like the behaviors identified in this paper, but have been less successful in effectively classifying finer scale behaviors associated with prey capture, prey handling and self-maintenance (Hammond et al., 2016; Ladds et al., 2017; Shamoun-Baranes et al., 2012). An initial partitioning into general behavior classes may simplify the process of defining detailed behavior profiles, especially where these behaviors occur as a subset within more coarse-scale behavior. While our results show that accurate classification of basic seabird behaviors can be developed using simple methods and a small group of accelerometer statistics, identifying fine scale behavior may require independently collected training data, and a larger suite of predictor variables, to capture the unique characteristics of less common behaviors.

5 | CONCLUSION

Obtaining reliable activity budgets from free-ranging animals is important for addressing a wide range of questions in wildlife ecology and animal behavior. Combined with methods for tracking animal location, behavioral classification from accelerometers could be used to examine the relationship between behavior and environmental conditions over large spatial and temporal scales. We believe that uncertainty about how to classify behavior from accelerometers has been a barrier to wider use of this technique. Our results demonstrate that general behaviors of seabirds can be classified from acceleration profiles using a range of techniques and a small number of predictor variables. Choice of classification method had a negligible effect on accuracy, therefore, researchers should not be impeded by a need to develop and apply the most advanced classification method, as multiple methods can provide similar results when classifying a small number of common behaviors. However, this finding may not hold in cases where the objective is to identify more detailed types of behavior than the broad classes considered here. Where the goal of classification is to develop a daily activity budget or estimate DEE, then simple classification methods are likely adequate, at least for waterbirds that primarily use flapping flight. Where the goal is to examine how different factors effect behavior, the HMM approach may be preferable because this approach can be used to directly test the effect of predictor variables on behavior.

ACKNOWLEDGEMENTS

R. Armstrong at the Nunavut Research Institute and M. Janssen, C. Geoffroy, and H. Hennin at National Wildlife Research Centre in Ottawa provided logistical support for research on murres. J. Nakoolak, J. Angootealuk, S. Richard, A. Eby, and S. Poole provided assistance in the field. We thank A. Nowlin for help with tagging, and Erica L. and J. Aragon for general help and support with field work on kittiwakes. J. Welcker generously loaned accelerometers for kittiwake research. Research on murres was undertaken with approval of the Animal Care Committee at McGill University (2015-7599) with funding from the National Sciences and Engineering Research Council of Canada, the Northern Contaminants Program from Indian and Northern Affairs Canada and the Newfoundland and Labrador Murre Fund with support from Wildlife Habitat Canada and Bird Studies Canada. Research on kittiwakes was undertaken in collaboration with the Institute for Seabird Research and Conservation with approval of the Animal Care Committee at the University of Manitoba (permit no. F11-020) and funding from the British Ornithological Union. A. Patterson received support from NSERC CREATE Environmental Innovation and the Garfield Weston Foundation. We also thank two anonymous reviewers for their comments, which greatly improved this manuscript.

CONFLICT OF INTEREST

None declared.

DATA ACCESSIBILITY

Data used in this analysis and R scripts for behavioral classifications have been archived at https://datadryad.org/(https://doi.org/10.5061/dryad.2hf101c).

ORCID

Allison Patterson https://orcid.org/0000-0001-9931-2693

REFERENCES

Angel, L., Berlincourt, M., & Arnould, J. (2016). Pronounced inter-colony variation in the foraging ecology of Australasian gannets: Influence of habitat differences. *Marine Ecology Progress Series*, *556*, 261–272. https://doi.org/10.3354/meps11845

Berlincourt, M., Angel, L. P., & Arnould, J. P. Y. (2015). Combined use of GPS and accelerometry reveals fine scale three-dimensional foraging behaviour in the short-tailed shearwater. *PLoS ONE*, *10*(10), e0139351. https://doi.org/10.1371/journal.pone.0139351

Bidder, O. R., Campbell, H. A., Gómez-Laich, A., Urgé, P., Walker, J., Cai, Y., ... Wilson, R. P. (2014). Love thy neighbour: Automatic animal behavioural classification of acceleration data using the K-nearest neighbour algorithm. *PLoS ONE*, *9*(2), e88609. https://doi.org/10.1371/journal.pone.0088609

Birt-Friesen, V. L., Montevecchi, W. A., Cairns, D. K., & Macko, S. A. (1989). Activity-specific metabolic rates of free-living northern gannets and other seabirds. *Ecology*, *70*(2), 357–367. https://doi.org/10.2307/1937540

Bishop, C. M., Spivey, R. J., Hawkes, L. A., Batbayar, N., Chua, B., Frappell, P. B., ... Butler, P. J. (2015). The roller coaster flight strategy of bar-headed geese conserves energy during Himalayan migrations. *Science*, *347*(6219), 250–254. https://doi.org/10.1126/science.1258732

Bom, R. A., Bouten, W., Piersma, T., Oosterbeek, K., & van Gils, J. A. (2014). Optimizing acceleration-based ethograms: The use of variable-time versus fixed-time segmentation. *Movement Ecology*, *2*, 6. https://doi.org/10.1186/2051-3933-2-6

Brown, D. D., Kays, R., Wikelski, M., Wilson, R., & Klimley, A. P. (2013). Observing the unwatchable through acceleration logging of animal behavior. *Animal Biotelemetry*, *1*, 20. https://doi.org/10.1186/2050-3385-1-20

Cagnacci, F., Boitani, L., Powell, R. A., & Boyce, M. S. (2010). Animal ecology meets GPS-based radiotelemetry: A perfect storm of opportunities and challenges. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *365*(1550), 2157–2162. https://doi.org/10.1098/rstb.2010.0107

Chimienti, M., Cornulier, T., Owen, E., Bolton, M., Davies, I. M., Travis, J. M. J., & Scott, B. E. (2016). The use of an unsupervised learning approach for characterizing latent behaviors in accelerometer data. *Ecology and Evolution*, *6*(3), 727–741. https://doi.org/10.1002/ece3.1914

Chivers, L. S., Hatch, S. A., & Elliott, K. H. (2016). Accelerometry reveals an impact of short-term tagging on seabird activity budgets. *The Condor*, *118*(1), 159–168. https://doi.org/10.1650/CONDOR-15-66.1

Collins, P. M., Green, J. A., Warwick-Evans, V., Dodd, S., Shaw, P. J. A., Arnould, J. P. Y., & Halsey, L. G. (2015). Interpreting behaviors from accelerometry: A method combining simplicity and objectivity. *Ecology and Evolution*, *5*(20), 4642–4654. https://doi.org/10.1002/ece3.1660

Elliott, K. H., Davoren, G. K., & Gaston, A. J. (2007). The influence of buoyancy and drag on the dive behaviour of an Arctic seabird, the Thick-billed Murre. *Canadian Journal of Zoology*, *85*(3), 352–361. https://doi.org/10.1139/Z07-012

Elliott, K. H., Ricklefs, R. E., Gaston, A. J., Hatch, S. A., Speakman, J. R., & Davoren, G. K. (2013). High flight costs, but low dive costs, in auks support the biomechanical hypothesis for flightlessness in penguins. *Proceedings of the National Academy of Sciences*, *110*(23), 9380–9384. https://doi.org/10.1073/pnas.1304838110

Elliott, K. H., Chivers, L. S., Bessey, L., Gaston, A. J., Hatch, S. A., Kato, A., ... Hare, J. F. (2014). Windscapes shape seabird instantaneous energy costs but adult behavior buffers impact on offspring. *Movement Ecology*, *2*, 17. https://doi.org/10.1186/s40462-014-0017-2

Elliott, K. H., Le Vaillant, M., Kato, A., Gaston, A. J., Ropert-Coudert, Y., Hare, J. F., ... Croll, D. (2014). Age-related variation in energy expenditure in a long-lived bird within the envelope of an energy ceiling. *Journal of Animal Ecology*, *83*(1), 136–146. https://doi.org/10.1111/1365-2656.12126

Furness, R. W., & Camphuysen, K. (C. J.). (1997). Seabirds as monitors of the marine environment. *ICES Journal of Marine Science*, *54*(4), 726–737. https://doi.org/10.1006/jmsc.1997.0243

Gómez Laich, A., Wilson, R. P., Gleiss, A. C., Shepard, E. L. C., & Quintana, F. (2011). Use of overall dynamic body acceleration for estimating energy expenditure in cormorants: Does locomotion in different media affect relationships? *Journal of Experimental Marine Biology and Ecology*, *399*(2), 151–155. https://doi.org/10.1016/j.jembe.2011.01.008

Grémillet, D., Dell'Omo, G., Ryan, P. G., Peters, G., Ropert-Coudert, Y., & Weeks, S. J. (2004). Offshore diplomacy, or how seabirds mitigate intra-specific competition: A case study based on GPS tracking of Cape gannets from neighbouring colonies. *Marine Ecology Progress Series*, *268*, 265–279. https://doi.org/10.3354/meps268265

Hammond, T. T., Springthorpe, D., Walsh, R. E., & Berg-Kirkpatrick, T. (2016). Using accelerometers to remotely and automatically characterize behavior in small animals. *Journal of Experimental Biology*, *219*(11), 1618–1624. https://doi.org/10.1242/jeb.136135

Hebblewhite, M., & Haydon, D. T. (2010). Distinguishing technology from biology: A critical review of the use of GPS telemetry data in ecology. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *365*(1550), 2303–2312. https://doi.org/10.1098/rstb.2010.0087

Jodice, P. G. R., Roby, D. D., Suryan, R. M., Irons, D. B., Kaufman, A. M., Turco, K. R., & Visser, G. H. (2003). Variation in energy expenditure among black-legged kittiwakes: Effects of activity-specific metabolic rates and activity budgets. *Physiological and Biochemical Zoology*, *76*(3), 375–388. https://doi.org/10.1086/375431

Jodice, P. G., Roby, D. D., Suryan, R. M., Irons, D. B., Turco, K. R., Brown, E. D., ... Visser, G. H. (2006). Increased energy expenditure by a seabird in response to higher food abundance. *Marine Ecology Progress Series*, *306*, 283–293. https://doi.org/10.3354/meps306283

Kuhn, M. (2016). A short introduction to the caret package. Retrieved from http://camoruco.ing.uc.edu.ve/cran/web/packages/caret/vignettes/caret.pdf

Ladds, M. A., Thompson, A. P., Kadar, J.-P., J Slip, D., P Hocking, D., & G Harcourt, R. (2017). Super machine learning: Improving accuracy and reducing variance of behaviour classification from accelerometry. *Animal Biotelemetry*, *5*, 8. https://doi.org/10.1186/s40317-017-0123-1

Langrognet, F., Lebret, R., Poli, C., & Iovleff, S. (2016). Rmixmod: supersed, unsupervised, semi-supervised classification with MIXture MODelling (Interface of MIXMOD Software). R Package Version 2.1.1. Retrieved from https://CRAN.R-project.org/package=Rmixmod

Lenth, R. V. (2016). Least-Squares Means: The R Package lsmeans. *Journal of Statistical Software*, 069(i01), Retrieved from http://econpapers.repec.org/article/jssjstsof/v_3a069_3ai01.htm

Leos-Barajas, V., Gangloff, E. J., Adam, T., Langrock, R., van Beest, F. M., Nabe-Nielsen, J., & Morales, J. M. (2017). Multi-scale modeling of animal movement and general behavior data using hidden Markov models with hierarchical structures. *Journal of Agricultural, Biological and Environmental Statistics*, *22*(3), 232–248. https://doi.org/10.1007/s13253-017-0282-9

Leos-Barajas, V., Photopoulou, T., Langrock, R., Patterson, T. A., Watanabe, Y. Y., Murgatroyd, M., & Papastamatiou, Y. P. (2016). Analysis of animal accelerometer data using hidden Markov models. *Methods in Ecology and Evolution*, *8*(2), 161–173. https://doi.org/10.1111/2041-210X.12657

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

MacArthur, R. A., Geist, V., & Johnston, R. H. (1982). Cardiac and behavioral responses of mountain sheep to human disturbance. *The Journal of Wildlife Management*, *46*(2), 351–358. https://doi.org/10.2307/3808646

McClintock, B. T., & Michelot, T. (2018). momentuHMM: R package for generalized hidden Markov models of animal movement. *Methods in Ecology and Evolution*, *9*(6), 1518–1530. https://doi.org/10.1111/2041-210X.12995

Mendez, L., Borsa, P., Cruz, S., de Grissac, S., Hennicke, J., Lallemand, J., … Weimerskirch, H. (2017). Geographical variation in the foraging behaviour of the pantropical red-footed booby. *Marine Ecology Progress Series*, *568*, 217–230. https://doi.org/10.3354/meps12052

Moreau, M., Siebert, S., Buerkert, A., & Schlecht, E. (2009). Use of a triaxial accelerometer for automated recording and classification of goats' grazing behaviour. *Applied Animal Behaviour Science*, *119*(3), 158–170. https://doi.org/10.1016/j.applanim.2009.04.008

Nathan, R., Spiegel, O., Fortmann-Roe, S., Harel, R., Wikelski, M., & 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. https://doi.org/10.1242/jeb.058602

Pagano, A., Rode, K., Cutting, A., Owen, M., Jensen, S., Ware, J., … Williams, T. (2017). Using tri-axial accelerometers to identify wild polar bear behaviors. *Endangered Species Research*, *32*, 19–33. https://doi.org/10.3354/esr00779

Pennycuick, C. J. (1987). Flight of auks (Alcidae) and other northern seabirds compared with southern Procellariiformes: Ornithodolite observations. *Journal of Experimental Biology*, *128*(1), 335–347.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core, Team (2018). *nlme: Linear and nonlinear mixed effects models*. R package version 3.1-137. Retrieved from https://CRAN.R-project.org/package=nlme

Quiros, A. L. (2007). Tourist compliance to a Code of Conduct and the resulting effects on whale shark (*Rhincodon typus*) behavior in Donsol, Philippines. *Fisheries Research*, *84*(1), 102–108. https://doi.org/10.1016/j.fishres.2006.11.017

Resheff, Y. S., Rotics, S., Harel, R., Spiegel, O., & Nathan, R. (2014). AcceleRater: A web application for supervised learning of behavioral modes from acceleration measurements. *Movement Ecology*, *2*, 27. https://doi.org/10.1186/s40462-014-0027-0

Robson, A. A., Chauvaud, L., Wilson, R. P., & Halsey, L. G. (2012). Small actions, big costs: The behavioural energetics of a commercially important invertebrate. *Journal of the Royal Society Interface*, *9*(72), 1486–1498. https://doi.org/10.1098/rsif.2011.0713

Ropert-Coudert, Y., Grémillet, D., Kato, A., Ryan, P. G., Naito, Y., & Le Maho, Y. (2004). A fine-scale time budget of Cape gannets provides insights into the foraging strategies of coastal seabirds. *Animal Behaviour*, *67*(5), 985–992. https://doi.org/10.1016/j.anbehav.2003.09.010

Ryan, P. G., Petersen, S. L., Peters, G., & Grémillet, D. (2004). GPS tracking a marine predator: The effects of precision, resolution and sampling rate on foraging tracks of African Penguins. *Marine Biology*, *145*(2), 215–223. https://doi.org/10.1007/s00227-004-1328-4

Sakamoto, K. Q., Sato, K., Ishizuka, M., Watanuki, Y., Takahashi, A., Daunt, F., & Wanless, S. (2009). Can ethograms be automatically generated using body acceleration data from free-ranging birds? *PLoS ONE*, *4*(4), e5379. https://doi.org/10.1371/journal.pone.0005379

Sato, K., Daunt, F., Watanuki, Y., Takahashi, A., & Wanless, S. (2008). A new method to quantify prey acquisition in diving seabirds using wing stroke frequency. *Journal of Experimental Biology*, *211*(1), 58–65. https://doi.org/10.1242/jeb.009811

Sato, N., Kokubun, N., Yamamoto, T., Watanuki, Y., Kitaysky, A. S., & Takahashi, A. (2015). The jellyfish buffet: Jellyfish enhance seabird foraging opportunities by concentrating prey. *Biology Letters*, *11*(8), 20150358. https://doi.org/10.1098/rsbl.2015.0358

Shamoun-Baranes, J., Bom, R., van Loon, E. E., Ens, B. J., Oosterbeek, K., & Bouten, W. (2012). From sensor data to animal behaviour: An oystercatcher example. *PLoS ONE*, *7*(5), e37997. https://doi.org/10.1371/journal.pone.0037997

Shepard, E., Wilson, R., Halsey, L., Quintana, F., Gómez Laich, A., Gleiss, A., … Norman, B. (2008). Derivation of body motion via appropriate smoothing of acceleration data. *Aquatic Biology*, *4*, 235–241. https://doi.org/10.3354/ab00104

Shepard, E., Wilson, R., Quintana, F., Gómez Laich, A., Liebsch, N., Albareda, D., … McDonald, D. (2008). Identification of animal movement patterns using tri-axial accelerometry. *Endangered Species Research*, *10*, 47–60. https://doi.org/10.3354/esr00084

Speakman, J. (1997). *Doubly labelled water: Theory and practice*. New York, NY: Chapman and Hall USA.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York, NY: Springer Science and Business Media.

Wakefield, E., Phillips, R., & Matthiopoulos, J. (2009). Quantifying habitat use and preferences of pelagic seabirds using individual movement data: A review. *Marine Ecology Progress Series*, *391*, 165–182. https://doi.org/10.3354/meps08203

Watanuki, Y., Niizuma, Y., Geir, W. G., Sato, K., & Naito, Y. (2003). Stroke and glide of wing-propelled divers: Deep diving seabirds adjust surge frequency to buoyancy change with depth. *Proceedings of the Royal Society B: Biological Sciences*, *270*(1514), 483–488. https://doi.org/10.1098/rspb.2002.2252

Weimerskirch, H., Le Corre, M., & Bost, C. (2008). Foraging strategy of masked boobies from the largest colony in the world: Relationship to environmental conditions and fisheries. *Marine Ecology Progress Series*, *362*, 291–302. https://doi.org/10.3354/meps07424

Wiemerskirch, H., Bishop, C., Jeanniard-du-Dot, T., Prudor, A., & Sachs, G. (2016). Frigate birds track atmospheric conditions over months-long transoceanic flights, *353*. Retrieved from http://science.sciencemag.org/content/sci/353/6294/74.full.pdf

Wilmers, C. C., Nickel, B., Bryce, C. M., Smith, J. A., Wheat, R. E., & Yovovich, V. (2015). The golden age of bio-logging: How animal-borne sensors are advancing the frontiers of ecology. *Ecology*, *96*(7), 1741–1753. https://doi.org/10.1890/14-1401.1

Wilson, R. P., Holton, M. D., Walker, J. S., Shepard, E. L. C., Scantlebury, D. M., Wilson, V. L., … Jones, M. W. (2016). A spherical-plot solution to linking acceleration metrics with animal performance, state, behaviour and lifestyle. *Movement Ecology*, *4*, 22. https://doi.org/10.1186/s40462-016-0088-3

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Patterson A, Gilchrist HG, Chivers L, Hatch S, Elliott K. A comparison of techniques for classifying behavior from accelerometers for two species of seabird. *Ecol Evol*. 2019;9:3030–3045. https://doi.org/10.1002/ece3.4740