The Canadian Journal of Statistics
Vol. 50, No. 1, 2022, Pages 327–356
La revue canadienne de statistique

Modelling multi-scale, state-switching functional data with hidden Markov models

Evan SIDROW[1*], Nancy HECKMAN[1], Sarah M. E. FORTUNE[2], Andrew W. TRITES[3,4], Ian MURPHY[5], and Marie AUGER-MÉTHÉ[1,4]

[1]*Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4*
[2]*Marine Mammal Research Unit, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4*
[3]*Department of Zoology, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4*
[4]*Institute for the Oceans and Fisheries, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4*
[5]*Department of Biostatistics, University of Florida, Gainesville, FL 32611, U.S.A.*

Key words and phrases: Accelerometer data; animal movement; biologging; diving behaviour; hierarchical modelling; killer whales; state switching; statistical ecology; time series.

MSC 2020: Primary 62M05; secondary 62P12.

Abstract: Data sets composed of sequences of curves sampled at high frequencies in time are increasingly common in practice, but they can exhibit complicated dependence structures that cannot be modelled using common methods in functional data analysis. We detail a hierarchical approach that treats the curves as observations from a hidden Markov model. The distribution of each curve is then defined by another fine-scale model that may involve autoregression and require data transformations using moving-window summary statistics or Fourier analysis. This approach is broadly applicable to sequences of curves exhibiting intricate dependence structures. As a case study, we use this framework to model the fine-scale kinematic movements of a northern resident killer whale (*Orcinus orca*) off the western coast of Canada. Through simulations, we show that our model produces more interpretable state estimation and more accurate parameter estimates compared to existing methods. *The Canadian Journal of Statistics* 50: 327–356; 2022 © 2021 Statistical Society of Canada

Résumé: Il est de plus en plus courant de rencontrer en pratique des ensembles de données composés de séquences de courbes échantillonnées à des fréquences élevées dans le temps. Ce type de données présentent la difficulté d'avoir des structures de dépendance complexes qui ne peuvent pas être modélisées à l'aide de méthodes de l'analyse des données fonctionnelles usuelles. Les auteurs de ce travail présentent une approche hiérarchique qui traite ces courbes comme des observations à partir d'un modèle Markov caché. La distribution de chaque courbe est ensuite définie par un autre modèle à échelle fine qui peut faire appel à des fonctions d'autorégression et utiliser des transformations de données de type résumés numériques à fenêtre mobile ou encore une analyse de Fourier. Cette approche est largement applicable aux séquences de courbes présentant des structures de dépendance complexes. Sa mise en pratique est illustrée à travers la construction d'un modèle à échelle fine des mouvements cinématiques des Epaulards (Orcinus orca) résidants de la zone nordique au large de la côte ouest du Canada. Grâce à des simulations, les auteurs réussissent à montrer que le modèle proposé produit de meilleures estimations d'état et de paramètres comparativement aux méthodes existantes et ce tant au niveau de la précision que de l'interprétation. *La revue canadienne de statistique* 50: 327–356; 2022 © 2021 Société statistique du Canada

Additional Supporting Information may be found in the online version of this article at the publisher's website.

* *Corresponding author: evan.sidrow@stat.ubc.ca*

1. INTRODUCTION

Biologging technology now provides researchers with kinematic data collected almost continuously in time (Hooten, King & Langrock, 2017). The collection and analysis of data from devices such as accelerometers have brought new insights to research tasks ranging from monitoring machine health (Getman et al., 2009) to understanding physical activity levels in children (Morris et al., 2006). The study of animal movement in particular has been transformed by tracking devices that record kinematic information in a variety of environments (Jeanniard-du Dot et al., 2016; Börger et al., 2020). Tags can record over 50 observations per second, resulting in time series containing millions of observations over the course of several hours. These data contain a wealth of information about human and animal behaviour, but modelling these large data sets poses a challenge for statisticians and biologists. One particular difficulty is that simultaneous coarse- and fine-scale processes are often reflected in high-frequency data, and each scale can exhibit a unique and complicated structure.

Biologging data are frequently modelled as a set of curves and analyzed using methods for functional data analysis, or FDA (e.g., Ramsay and Silverman, 2005). For example, Morris et al. (2006) views a child's activity level as a set of daily curves where metabolic activity is a function of time. Similarly, Fu & Heckman (2019) views the dive profile of a southern elephant seal (*Mirounga leonina*) as a set of dive curves. Dive amplitude and phase variation are used for classification into "dive types."

FDA was originally developed to process curves assumed to be independent replicates (i.e., there is no between-curve dependence). Within-curve, fine-scale structure is not usually incorporated in FDA models. However, sets of curves often exhibit complex sequential dependencies, both between and within curves. This is especially the case for biologging data (Leos-Barajas et al., 2017). On a coarse scale, the dive profiles of marine animals show a discrete number of distinct dive types, where the sequence of dive types exhibits state-switching behaviour (Tennessen et al., 2019a). On a fine scale, these profiles can display bouts of short-term periodicity within each dive (Adam et al., 2019). Fine-scale periodicity nested within a coarser state-switching process is also common in fields such as machine health (Xin, Hamzaoui & Antoni, 2018; Lucero et al., 2019) and speech recognition (Juang & Rabiner, 1991).

Some FDA models account for between-curve dependence that occurs when multiple curves arise from separate groups of individuals, but these models are inadequate when modelling certain kinds of time dependencies. For example, previous studies have used multi-level models with random effects to model variation between and within individuals in the daily activity levels of children (Morris et al., 2006) or in the menstrual cycles of adults (Brumback & Rice, 1998). More recent work involving multi-level models includes that of Crainiceanu, Staicu & Di (2009), Di et al. (2009), and Chen & Müller (2012). However, the first two papers do not account for temporal between-curve dependence and the third article's model of temporal dependence assumes that curves evolve smoothly in time. This is not appropriate in many biologging applications, where behaviours often change suddenly between a discrete number of types. In addition to multi-level models, FDA researchers have used functional time series to model dependence in a sequence of curves. Functional time series extends the ideas of classic time series to model the evolution of one curve into the next (Kokoszka & Reimherr, 2018), but does not account for sequences of time-series curves whose distributions are determined by a discrete number of well-defined hidden states.

Traditional FDA techniques similarly fall short when modelling complicated within-curve data. In particular, within-curve structure is usually modelled by a generic smooth mean function and a covariance function (Yao, Müller & Wang, 2005) or with random regression (Rice & Wu, 2001). However, time-series data exhibiting both sharp behavioural changes and periodic fine-scale structure are difficult to model with these classical FDA techniques.

 DOI: 10.1002/cjs.11673

Our goal is to identify and describe discrete behavioural states at multiple scales within functional data (e.g., coarse-scale curve types and fine-scale behavioural states) and to model the dependence structure between those states. To accomplish this, we turn to the field of animal movement modelling (Hooten, King & Langrock, 2017), where one of the most prevalent techniques of late is the hidden Markov model, or HMM (Patterson et al., 2017; McClintock et al., 2020). HMMs interpret animal movement data as arising from a Markov chain on a discrete number of behavioural states, allowing biologists to infer the underlying behaviour of an animal from sequential observations of its position. While ubiquitous in ecology literature, HMMs have seen little use in nonparametric functional modelling, with a few notable exceptions. In particular, Langrock et al. (2018) takes a nonparametric approach to model the distributions of HMM observations with B-splines, while de Souza & Heckman (2014) and de Souza, Heckman & Xu (2017) use HMMs to model state-switching behaviour within functional data. However, none of these papers accounts for temporal correlation taking place on multiple scales.

While useful, HMMs alone are also not sufficient to model high-frequency time-series data for three primary reasons. First, classical HMMs fail to model simultaneous behavioural processes that occur at different time scales (e.g., both between and within curves). To address this issue, statistical ecologists have employed hierarchical HMMs (HHMMs) (Leos-Barajas et al., 2017; Adam et al., 2019), which model both scales with conditionally dependent HMMs. Second, HMMs assume that subsequent observations are independent given an underlying hidden state process, but this is often not the case when observations are taken at extremely high frequencies. Several solutions have been proposed in the ecology literature, including the hidden movement Markov model (Whoriskey et al., 2016) and the conditionally autoregressive HMM, or CarHMM (Lawler et al., 2019). Third, traditional HMMs, CarHMMs, and HHMMs cannot easily capture complicated dependence structures on short time scales. For example, Adam et al. (2019) did not capture the fine-scale periodic swimming patterns of horn sharks (*Heterodontus francisci*) using a traditional HHMM. Heerah et al. (2017) successfully used Fourier analysis within an HMM to account for daily behavioural cycles in marine mammals. Fourier analysis has previously been used with accelerometer data to explain animal behaviour (Fehlmann et al., 2017; Shorter et al., 2017). Thus, incorporating Fourier analysis into the structure of an HMM appears to be a promising approach to account for fine-scale periodic structures in biologging data.

We combine these existing methods from statistical ecology literature in novel ways to classify and describe state-switching behaviours in functional data while accounting for complex temporal dependence. The resulting methods make up a tool box that can be used to build arbitrarily complex hierarchical models to explain multi-scale functional and time-series data with intricate dependence structures. We begin in Section 2 by describing HMMs as well as two variants, CarHMMs and HHMMs, and discuss how summary statistics over moving windows can handle fine-scale dependence structures. We also show how these methods can be combined to analyze increasingly complex data. In Section 3, we fit several candidate models to data from a killer whale (*Orcinus orca*) from the threatened northern resident population off the coast of British Columbia, Canada. Section 4 details a simulation study based on these candidate models, and in Section 5 we discuss our results.

2. MODELS AND PARAMETER ESTIMATION

Consider a sequence of T curves, where curve t is characterized by a curve-level (or coarse-scale) observation Y_t as well as a sequence of T_t^* within-curve (or fine-scale) observations Y_t^*. Namely, $Y_t^* = \left\{Y_{t,1}^*, \ldots, Y_{t,T_t^*}^*\right\}$ is made up of fine-scale quantities derived from curve t and indexed by t^*. Both Y_t and Y_{t,t^*}^* can be either vectors or scalars. We call the sequence of coarse-scale observations $Y = \left\{Y_1, \ldots, Y_T\right\}$ and the collection of all fine-scale observations $Y^* = \left\{Y_1^*, \ldots, Y_T^*\right\}$. We assume that the curves $Y_1, \ldots, Y_T$ are indexed according to the order in which they are observed in time,

but the T observations need not be equally spaced in time. To develop our model for this data, we detail the structure of a traditional HMM followed by three variations that generalize its base structure. We then show how each of these generalized HMMs can be synthesized to form a wide variety of more complicated models.

2.1. HMMs as a Base Structure

HMMs describe state-switching Markovian processes in discrete time and are the core structure we use to model both Y and Y^*. For simplicity, we focus on Y to introduce the model. An HMM is composed of a sequence of unobserved states $X = \{X_1, \ldots, X_T\}$ together with an observation sequence $Y = \{Y_1, \ldots, Y_T\}$, where X_t is associated with the observation Y_t. The Y_ts are often referred to as "emissions" and the index t typically refers to time. The X_ts form a Markov chain and can take integer values between 1 and N. Their distribution is governed by the distribution of the initial state X_1 and an $N \times N$ transition probability matrix Γ, where $\Gamma_{ij} = \Pr(X_{t+1} = j \mid X_t = i)$. We consider only time-homogeneous Markov chains, meaning that Γ does not depend on time. We assume that X_1 follows the chain's stationary distribution, which is denoted by an N-dimensional row vector δ, where $\delta_i = \Pr(X_1 = i)$. A Markov chain's stationary distribution is determined by its transition probability matrix via $\delta = \delta\Gamma$, where $\sum_{i=1}^{N} \delta_i = 1$. The distribution of an emission Y_t conditioned on the corresponding hidden state X_t does not depend upon any other observation or hidden state. If $X_t = i$, then we denote the conditional density or probability mass function of Y_t as $f^{(i)}(\cdot; \theta^{(i)})$ or simply $f^{(i)}(\cdot)$, where $\theta^{(i)}$ is a state-dependent parameter describing the emission distribution.

The joint likelihood of both the parameters and the hidden states is given by

$$\mathcal{L}_{\text{HMM}}(x, \theta, \Gamma; y) = \delta_{x_1} f^{(x_1)}\left(y_1; \theta^{(x_1)}\right) \prod_{t=2}^{T} \Gamma_{x_{t-1}x_t} f^{(x_t)}\left(y_t; \theta^{(x_t)}\right), \tag{1}$$

but it is also tractable to sum over all possible values of x to obtain

$$\mathcal{L}_{\text{HMM}}(\theta, \Gamma; y) = \delta P(y_1; \theta) \prod_{t=2}^{T} \Gamma P(y_t; \theta) \mathbf{1}_N, \tag{2}$$

where $\mathbf{1}_N$ is an N-dimensional column vector of 1's and $P(y_t; \theta)$ is an $N \times N$ diagonal matrix with the (i, i)th entry $f^{(i)}(y_t; \theta^{(i)})$. This expression can be evaluated in $\mathcal{O}(T)$ time using the well-known forward algorithm (Zucchini, Macdonald & Langrock, 2016).

We take a frequentist approach in this article, which involves obtaining maximum likelihood estimates $\{\hat{\theta}, \hat{\Gamma}\}$ from Equation (2). The estimated probability of each hidden state $\hat{\Pr}(X_t = i \mid Y = y)$ can then be calculated using $\{\hat{\theta}, \hat{\Gamma}\}$ and the forward–backward algorithm (Zucchini, Macdonald & Langrock, 2016). Alternatively, it is reasonable to use a Bayesian approach, where Equation (1) is combined with a prior distribution to obtain a posterior distribution over X, θ, and Γ. This posterior distribution can be sampled from using methods such as sequential Monte Carlo (Douc et al., 2011), Markov-chain Monte Carlo (Scott, 2002), or variational inference (Foti et al., 2014).

Following Leos-Barajas et al. (2017), we re-parameterize the $N \times N$ transition probability matrix Γ such that the entries of the matrix are forced to be non-negative and the rows sum to 1:

$$\Gamma_{ij} = \frac{\exp(\eta_{ij})}{\sum_{k=1}^{N} \exp(\eta_{ik})},$$

 DOI: 10.1002/cjs.11673

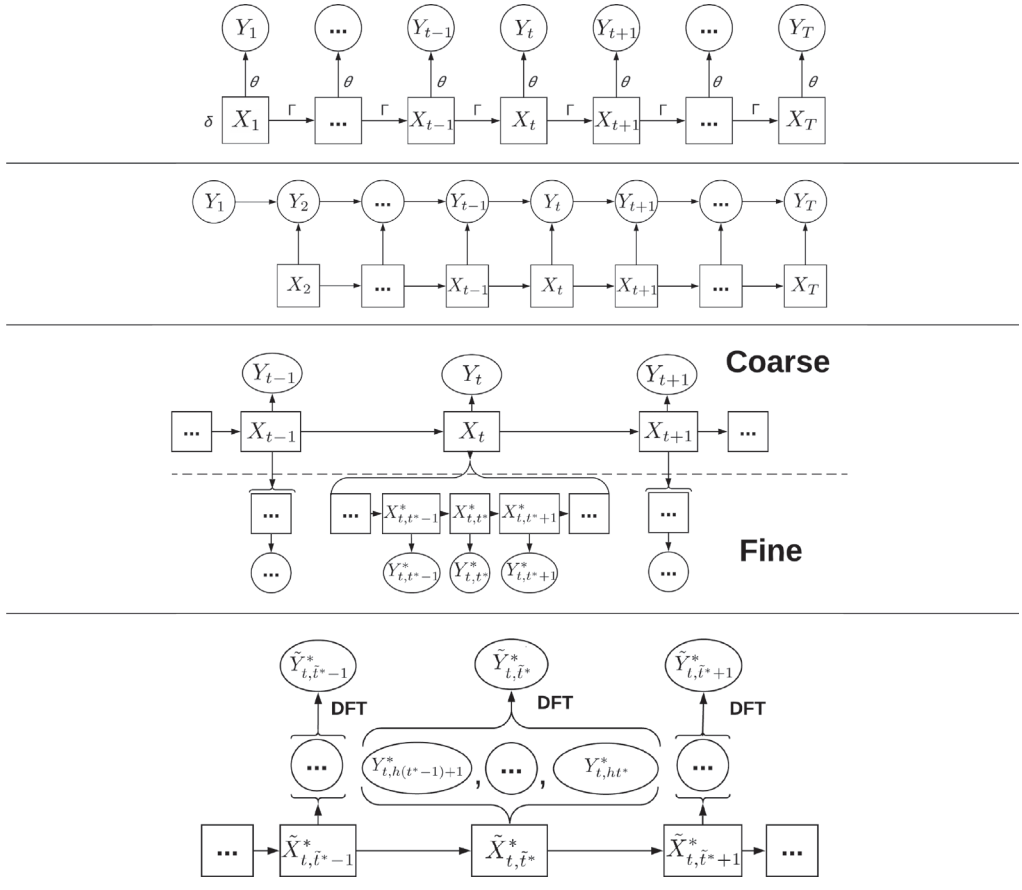

FIGURE 1: Dependence structure of a standard HMM (top), CarHMM (middle-top), HHMM (middle-bottom), and HMM-DFT (bottom). Hidden state sequences are denoted by X on the coarse scale and by X^* on the fine scale. Observations are denoted by Y on the coarse scale and by Y^* on the fine scale. The coarse-scale process is not included in the HMM-DFT because moving-window transformations often only apply to fine-scale data. The fine-scale observations of the HMM-DFT are transformed using a moving window and denoted by $\tilde{Y}^*$ with corresponding fine-scale hidden states $\tilde{X}^*$.

where $i, j = 1, \ldots, N$ and η_{ii} is set to zero for identifiability. This formulation simplifies likelihood maximization by removing constraints in the optimization problem. We assume there are no covariate effects, but similar to the work of de Souza, Heckman & Xu (2017) for independent states and Adam et al. (2019) for Markovian states, one could incorporate covariates into Γ by setting $\eta_{ij}(z_t) = \beta_{ij}^\top z_t$ for $i \neq j$, where z_t is a column vector of known covariates and β_{ij} is a column vector of unknown regression coefficients. For simplicity, we will continue to use Γ in our notation, suppressing the re-parameterization in terms of η. Figure 1 shows the dependence structure of an HMM.

2.2. Relaxing Conditional Independence with the CarHMM

The CarHMM (Lawler et al., 2019) is a generalization of the HMM, which explicitly models autocorrelation in the observation sequence beyond the correlation induced by the hidden state process. Like the traditional HMM, the CarHMM is made up of a Markov chain of unobserved

states $\{X_1, \ldots, X_T\}$, each taking integer values between 1 and N. The CarHMM also has a transition probability matrix Γ and initial distribution δ equal to the stationary distribution of Γ. Unlike the traditional HMM, the CarHMM assumes that the distribution of Y_t conditioned on $\{X_1, \ldots, X_T\}$ and $\{Y_1, \ldots, Y_{t-1}\}$ depends on both X_t and Y_{t-1} rather than only X_t. The first emission Y_1 is treated as a fixed initial value that does not depend upon X_1. We denote the conditional density or probability mass function of Y_t, given $Y_{t-1} = y_{t-1}$ and $X_t = i$, by $f^{(i)}(\cdot \mid y_{t-1}; \theta^{(i)})$, or simply $f^{(i)}(\cdot \mid y_{t-1})$. As a concrete example, if Y_t is a scalar, then one may assume that Y_t given $X_t = i$ is normally distributed with parameters $\theta^{(i)} = \left\{\mu^{(i)}, \sigma^{(i)}, \phi^{(i)}\right\}$, where

$$\mathbb{E}\left(Y_t \mid Y_{t-1} = y_{t-1}, X_t = i\right) = \phi^{(i)}\, y_{t-1} \; + \; \left(1 - \phi^{(i)}\right)\; \mu^{(i)}, \tag{3}$$

and

$$\mathbb{V}\left(Y_t \mid Y_{t-1} = y_{t-1}, X_t = i\right) = \left(\sigma^{(i)}\right)^2. \tag{4}$$

A CarHMM following Equations (3) and (4) can be viewed as a discrete-time version of a state-switching Ornstein–Uhlenbeck process (Michelot & Blackwell, 2019). This follows in the same way that an AR(1) process is the discrete-time version of a traditional Ornstein–Uhlenbeck process.

As above, the likelihood of θ and Γ corresponding to the CarHMM can be calculated using the forward algorithm. If y is the sequence of observations, then

$$\mathcal{L}_{\text{CarHMM}}(\theta, \Gamma; y) = \delta \prod_{t=2}^{T} \Gamma P\left(y_t \mid y_{t-1}; \theta\right) \mathbf{1}_N,$$

where $P(y_t \mid y_{t-1}; \theta)$ is an $N \times N$ diagonal matrix with the (i,i)th entry equal to $f^{(i)}\left(y_t \mid y_{t-1}; \theta^{(i)}\right)$. Figure 1 shows a graphical representation of the dependence structure of the CarHMM.

2.3. Incorporating Multiple Scales with the HHMM

An HHMM accounts for processes occurring simultaneously at different scales by modelling both the coarse-scale process and fine-scale process with either HMMs (Leos-Barajas et al., 2017; Adam et al., 2019) or CarHMMs, as defined in Sections 2.1 and 2.2. Recall that $\{X_1, \ldots, X_T\}$ is an unobserved Markov chain with N possible states, and $\{Y_1, \ldots, Y_T\}$ is the set of corresponding observations with state-dependent parameters $\theta^{(i)}$ for $i = 1, \ldots, N$. In the hierarchical setting, each state X_t also emits a sequence of fine-scale unobserved states, $X_t^* = \left\{X_{t,1}^*, \ldots, X_{t,T_t^*}^*\right\}$. In turn, X_t^* emits a sequence of fine-scale observations $Y_t^* = \left\{Y_{t,1}^*, \ldots, Y_{t,T_t^*}^*\right\}$. For each curve t, the parameters of the HMM (or CarHMM) describing the fine-scale process $\left\{X_t^*, Y_t^*\right\}$ depend on the value of X_t. If $X_t = i$, then X_{t,t^*}^* represents one of $N^{*(i)}$ possible fine-scale behaviours associated with the coarse-scale hidden state i. Furthermore, the components of X_t^* make up a Markov chain with an $N^{*(i)} \times N^{*(i)}$ transition probability matrix $\Gamma^{*(i)}$ and an initial distribution $\delta^{*(i)}$ which we assume is equal to the stationary distribution of the chain. The distribution of Y_{t,t^*}^*, given $Y_{t,t^*-1}^* = y_{t,t^*-1}^*$, $X_{t,t^*}^* = i^*$, and $X_t = i$, is governed by the parameter $\theta^{*(i,i^*)}$ and has a density or probability mass function denoted by $f^{*(i,i^*)}\left(\cdot \mid y_{t,t^*-1}^*; \theta^{*(i,i^*)}\right)$, or simply $f^{*(i,i^*)}\left(\cdot \mid y_{t,t^*-1}^*\right)$. We denote the set of fine-scale emission parameters corresponding to $X_t = i$ by $\theta^{*(i)} = \left\{\theta^{*(i,1)}, \ldots, \theta^{*\left(i,N^{*(i)}\right)}\right\}$. In summary,

1. $\{Y, X\}$ follows a (Car)HMM with $\Gamma \in \mathbb{R}^{N \times N}$,
2. $(Y_t \mid Y_{t-1} = y_{t-1}, X_t = i)$ has the density $f^{(i)}(\cdot \mid y_{t-1}; \theta^{(i)})$,

3. $\left\{Y_t^*, X_t^* \mid X_t = i\right\}$ follows a (Car)HMM with $\Gamma^{*(i)} \in \mathbb{R}^{N^{*(i)} \times N^{*(i)}}$, and
4. $\left(Y_{t,t^*}^* \mid Y_{t,t^*-1}^* = y_{t,t^*-1}^*, X_{t,t^*}^* = i^*, X_t = i\right)$ has the density $f^{*(i,i^*)}\left(\cdot \mid y_{t,t^*-1}^*; \theta^{(i,i^*)}\right)$.

Given the coarse-scale hidden state sequence X, the $T+1$ sets $\left\{X_1^*, Y_1^*\right\}, \ldots, \left\{X_T^*, Y_T^*\right\}$, and $\{Y_1, \ldots, Y_T\}$ are assumed to be independent of one another.

Forcing certain parameters to be shared can reduce the complexity and increase the interpretability of an HHMM. For example, since $f^{*(i,i^*)}$ depends on both $X_t = i$ and $X_{t,t^*}^* = i^*$, the distribution of Y_{t,t^*}^* is defined by the pair (X_t, X_{t,t^*}^*). However, in our killer whale case study (see Section 3), we take $N^{*(i)} = N^*$ for all i and share the fine-scale emission parameters across the N coarse-scale hidden states $\left(\text{i.e., } \theta^{*(1,i^*)} = \cdots = \theta^{*(N,i^*)} = \theta^{*(\cdot,i^*)} \text{ for all } i^* = 1, \ldots, N^*\right)$. As a result, the distribution of the fine-scale observation Y_{t,t^*}^* does not depend on X_t, but instead is defined exclusively by X_{t,t^*}^* $\left(\text{or } \left(\cdot, X_{t,t^*}^*\right) \text{ for notational consistency}\right)$.

Because of the nested structure of the HHMM, it is straightforward to extend the forward algorithm and sum over all possible coarse-scale hidden states X and all possible fine-scale hidden states X^* in the likelihood. Let $y = \{y_1, \ldots, y_T\}$ be the sequence of observed coarse-scale emissions and $y^* = \left\{y_1^*, \ldots, y_T^*\right\}$ be the collection of T observed fine-scale emission vectors. In addition, let $\theta^* = \left\{\theta^{*(1)}, \ldots, \theta^{*(N)}\right\}$ denote the collection of all fine-scale emission parameters, and let $\Gamma^* = \left\{\Gamma^{*(1)}, \ldots, \Gamma^{*(N)}\right\}$ denote the collection of all fine-scale transition probability matrices. The likelihood of the observed data is then

$$\mathcal{L}_{\text{HHMM}}\left(\theta, \theta^*, \Gamma, \Gamma^*; y, y^*\right) = \delta P_m\left(y_1, y_1^*; \theta, \theta^*, \Gamma^*\right) \prod_{t=2}^{T} \Gamma P_m\left(y_t, y_t^* \mid y_{t-1}; \theta, \theta^*, \Gamma^*\right) \mathbf{1}_N, \quad (5)$$

where $P_m(y_t, y_t^* \mid y_{t-1}; \theta, \theta^*, \Gamma^*)$ is an $N \times N$ diagonal matrix whose exact structure depends upon the coarse- and fine-scale models. If the coarse-scale model is an HMM, $P_m(y_1, y_1^*; \theta, \theta^*, \Gamma^*)$ and $P_m(y_t, y_t^* \mid y_{t-1}; \theta, \theta^*, \Gamma^*)$, for $t \geq 2$, both have (i,i)th entries equal to $f^{(i)}(y_t)\mathcal{L}_{\text{fine}}\left(\theta^{*(i)}, \Gamma^{*(i)}; y_t^*\right)$. If the coarse-scale model is a CarHMM, $P_m(y_1, y_1^*; \theta, \theta^*, \Gamma^*)$ has its (i,i)th entry equal to $\mathcal{L}_{\text{fine}}\left(\theta^{*(i)}, \Gamma^{*(i)}; y_1^*\right)$ and $P_m(y_t, y_t^* \mid y_{t-1}; \theta, \theta^*, \Gamma^*)$, for $t \geq 2$, has its (i,i)th entry equal to $f^{(i)}(y_t \mid y_{t-1})\mathcal{L}_{\text{fine}}\left(\theta^{*(i)}, \Gamma^{*(i)}; y_t^*\right)$. The fine-scale likelihood $\mathcal{L}_{\text{fine}}$ corresponds to the likelihood of the fine-scale model, which can be either a CarHMM or an HMM. Figure 1 displays the dependence structure of an HHMM.

2.4. Transforming Fine-scale Observations with the HMM-DFT

In many applications where data are collected at high frequencies, intricate dependency structures arise within the fine-scale process that cannot be adequately modelled with the HMM variations described thus far. To handle these additional fine-scale structures, we recommend replacing $Y_t^* = \left\{Y_{t,1}^*, \ldots, Y_{t,T_t^*}^*\right\}$ with relevant statistics that summarize any non-Markovian behaviour. To maintain the temporal structure of the fine-scale process, local summary statistics can be calculated from a moving window with the stride length h over the elements of Y_t^*. Stride length refers to the distance between the first element of consecutive windows, so a stride length of h implies that the first window starts at $Y_{t,1}^*$, the second at $Y_{t,1+h}^*$, and so on. Subject matter experts are often required to determine the specific summary statistics employed as well as the optimal window size and stride length of the moving window. Larger stride lengths result in a larger loss of information but also reduce the dimensionality of the fine-scale process, which allows for faster model fitting. In addition, setting the stride length equal to the window size avoids artificial residual correlation arising from overlapping windows.

We are interested in fine-scale summary statistics that are interpretable to practitioners and can effectively differentiate between fine-scale behavioural states. As a result, we use the discrete Fourier transform (DFT) of a moving forward window with a width of h and a stride of h across Y_t^*:

$$\text{DFT}\left\{Y_{t,t^*}^*, \ldots, Y_{t,t^*+h-1}^*\right\}(k) = \sum_{n=0}^{h-1} Y_{t,t^*+n}^* \exp\left(-\frac{i2\pi}{h}kn\right) \tag{6}$$

for $t^* = 1, h+1, 2h+1, \ldots$ and $k = 0, 1, \ldots, h-1$, where $i = \sqrt{-1}$. If Y_{t,t^*}^* is a vector, then the DFT is taken component-wise. We omit the final window if $t^* + h - 1$ exceeds T_t^*, denote the total number of windows by $\tilde{T}_t^*$, and index the windows with $\tilde{t}^* = 1, \ldots, \tilde{T}_t^*$. Next, we calculate the transformed observations $\tilde{Y}_{t,\tilde{t}^*}^* = \left\{\tilde{A}_{t,\tilde{t}^*}^*, \tilde{W}_{t,\tilde{t}^*}^*\right\}$ as

$$\tilde{A}_{t,\tilde{t}^*}^* = \frac{1}{h}\sum_{n=1}^{h} Y_{t,h(\tilde{t}^*-1)+n}^* \quad \text{and} \quad \tilde{W}_{t,\tilde{t}^*}^* = \sum_{k=1}^{\tilde{\omega}} \left|\left|\text{DFT}\left\{Y_{t,h(\tilde{t}^*-1)+1}^*, \ldots, Y_{t,h\tilde{t}^*}^*\right\}(k)\right|\right|_2^2, \tag{7}$$

where $\tilde{\omega} \leq h - 1$ is a problem-specific tuning parameter corresponding to the maximum recorded frequency within each window. In words, $\tilde{A}_{t,\tilde{t}^*}^*$ is the average value of Y_t^* within window $\tilde{t}^*$ and $\tilde{W}_{t,\tilde{t}^*}^*$ is the squared two-norm of the component of the window that can be attributed to frequencies between one and $\tilde{\omega}$ periods per window. More intuitively, $\tilde{W}_{t,\tilde{t}^*}^*$ corresponds to the "wiggliness" of the fine-scale data within curve t and window $\tilde{t}^*$.

After performing this transformation, the entire model must be redefined since the fine-scale HMM (or CarHMM) directly models the distribution of the summary statistics $\tilde{Y}_t^*$ rather than the discretized curve Y_t^*. Because $\tilde{Y}_t^*$ exists on a coarser scale than Y_t^*, there are only $\tilde{T}_t^*$ unobserved states associated with $\tilde{Y}_t^* = \left\{\tilde{Y}_{t,1}^*, \ldots, \tilde{Y}_{t,\tilde{T}_t^*}^*\right\}$. We denote these unobserved states as $\tilde{X}_t^* = \left\{\tilde{X}_{t,1}^*, \ldots, \tilde{X}_{t,\tilde{T}_t^*}^*\right\}$. The fine-scale transition probability matrices $\Gamma^{*(i)}$ and probability density functions $f^{*(i,i^*)}$ correspond to $\tilde{X}_t^*$ and $\tilde{Y}_t^*$ rather than X_t^* and Y_t^*, as do all fine-scale model assumptions (e.g., conditional independence, autoregressive structure between observations, etc.). Fine-scale model selection and validation should therefore be adjusted accordingly.

The likelihood of this model is identical to that of the original HMM or CarHMM defined in Sections 2.1 and 2.2, but Y^* is replaced with $\tilde{Y}^*$ and X^* is replaced with $\tilde{X}^*$. To clearly differentiate the models, we refer to an HMM with $\tilde{Y}^*$ as observations and $\tilde{X}^*$ as hidden states as an HMM-DFT. Figure 1 displays the dependence structure of a fine-scale HMM-DFT.

Replacing Y_t^* with summary statistics results in a loss of information since it involves substituting infinite-dimensional curves with a finite-dimensional representation. While we are primarily interested in classifying and describing coarse- and fine-scale behaviour, some researchers may be less interested in interpretability and more interested in predicting future functional data. To this end, Aue, Norinho & Hörmann (2015) and Gao, Shang & Yang (2019) both use functional principal component analysis (FPCA) and derive bounds on the reconstruction error caused by dimension reduction. While these papers do not account for state-switching behaviour between curves or windows, deriving similar error bounds for FPCA within an HMM framework appears to be a promising direction for future research.

2.5. Generalized Hierarchical Markov Models

Traditional HHMMs treat both the coarse-scale and the fine-scale processes as realizations of an HMM or CarHMM. However, the fine-scale observations of a particular dive Y_t^* can be modelled using a large variety of parametric models that admit easy-to-compute likelihoods or penalized

likelihoods. As such, the fine-scale HMM likelihood term $\mathcal{L}_{\text{fine}}$ in Equation (5) can be replaced by the likelihood of a general fine-scale model whose parameters depend upon the coarse-scale hidden state. For example, Bebbington (2007) and Borchers et al. (2013) investigated data sets with count onsets as observations, so they used variations of a Poisson process as their fine-scale model. If the fine-scale model is a simple Poisson process, then this approach is equivalent to a Markov-modulated Poisson process (Fischer & Meier-Hellstern, 1993). The fine-scale process can also be modelled similarly to Langrock et al. (2018), which uses B-splines to model the emission distribution of an HMM. This nonparametric approach uses a penalized likelihood term that can easily replace the usual fine-scale likelihood term in Equation (5). Another class of fine-scale models is the set of continuous time methods such as the continuous-time HMM (CTHMM) (Liu et al., 2015) and the state-switching Ornstein–Uhlenbeck process (Michelot & Blackwell, 2019). A CTHMM may be appropriate if observations are not equally spaced in time (Liu et al., 2015). Xu, Laber & Staicu (2020) modelled high-frequency biologging accelerometer data from individuals by incorporating a CTHMM into a hierarchical model similar to ours. However, they assumed that individuals are partitioned into subgroups a priori, whereas we use an HMM to infer coarse-scale hidden states.

These examples include a few of many fine-scale models that can act as initial building blocks in a practitioner's toolbox to construct increasingly complex hierarchical models. A myriad of possible models can be built using this framework, but these models can quickly become complicated and computationally expensive to fit. Therefore, models should be constructed with care to achieve an adequate fit to the data while avoiding overfitting and high computational costs.

3. KILLER WHALE CASE STUDY

To illustrate the process of constructing a model using these building blocks, we analyze the diving behaviour of a northern resident killer whale in Queen Charlotte Sound, off the coast of British Columbia, and construct several candidate models to categorize and describe its diving behaviour.

Understanding animal behaviour is important for conservation efforts, as environmental changes caused by anthropogenic activity can directly impact animal behaviour (Sutherland, 1998). HMMs have been used to understand how diving behaviours of various species are affected by disturbances (e.g., DeRuiter et al., 2017; Isojunno et al., 2017). For killer whales, we are interested in categorizing different diving behaviours and identifying potential foraging dives. Northern resident killer whales feed almost exclusively on calorie-rich Chinook salmon (*Oncorhynchus tshawytscha*) (Ford & Ellis, 2006), which typically occur deeper and are less numerous than smaller types of salmon (Ford et al., 2009). Northern resident killer whales therefore must expend significant amounts of energy to capture Chinook salmon (Williams & Noren, 2009; Noren, 2011; Wright et al., 2017). Acceleration data can be used to estimate an animal's energy expenditure (Green et al., 2009; Wilson et al., 2019), but the animal's behavioural state must be accounted for in order to obtain accurate estimates (Jeanniard du Dot et al., 2016). Therefore, understanding both the behavioural state of the killer whale and the distribution of acceleration within each behavioural state is needed to determine the true energetic requirements of the animal.

3.1. Data Collection and Preprocessing

The data we use were collected on 2 September 2019 from 12:49 PM to 6:06 PM PDT, and consist of depth and acceleration over time. Observations were collected at a rate of 50 Hz using a CATS biologger (Customizable Animal Tracking Solutions, www.cats.is). Acceleration was measured in three dimensions which, together, represent the complete range of movement of

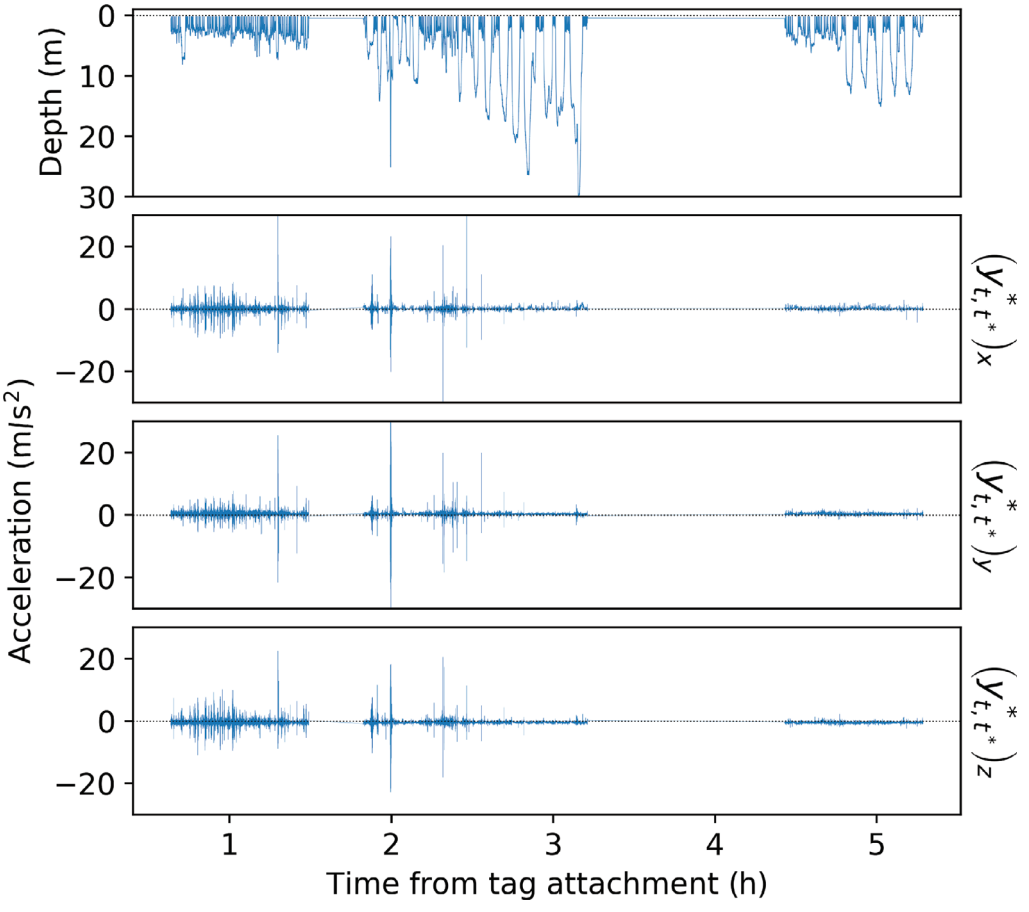

FIGURE 2: Dive depth (top panel) and three-dimensional acceleration (bottom three panels) from a killer whale over approximately 5 h. An exact physical interpretation of each component of acceleration is difficult due to variations in tag orientation. There are data gaps occurring from around hours 1.5 to 1.8 and from around hours 3.2 to 4.5. Both data gaps are excluded from analyses.

an animal (forward/backward, upward/downward, and right/left). Tri-axial acceleration readings are common in these types of tags and are often used to infer animal behaviour such as foraging (Fehlmann et al., 2017; Wright et al., 2017; Cade et al., 2018). The act of attaching and detaching the tag caused anomalous behaviour before 1:20 PM and after 6:00 PM, so observations taken during these time periods are ignored. There were also periods of time when the tag failed to record observations, resulting in data gaps between 2:25 PM and 2:37 PM and between 4:07 PM and 5:07 PM. To preprocess the data, we smooth the depth and acceleration curves by taking a moving average within a window of 1/10 of a second. We then define a killer whale "dive" as any continuous interval of data that occurs below 0.5 m in depth and lasts for at least 10 s. Data are preprocessed in part with the *divebomb* package in Python (Nunes, 2019). The preprocessed data contain a total of 267 dives, all of which are displayed in Figure 2. Each dive is treated as one curve, and the sequence of dives makes up the coarse-scale process. Specifically, the observed coarse-scale observations $y = \{y_1, \ldots, y_{267}\}$ make up a sequence of dive durations in seconds, and the coarse-scale hidden states $x = \{x_1, \ldots, x_{267}\}$ represent the corresponding dive

types. The fine-scale observations for dive t are measured in units of m/s^2 and are contained in $y_t^* = \left\{y_{t,1}^*, \ldots, y_{t,T_t^*}^*\right\}$. The fine-scale hidden states of dive t, $x_t^* = \left\{x_{t,1}^*, \ldots, x_{t,T_t^*}^*\right\}$, represent the subdive behavioural states of the killer whale. The collection of all acceleration data is denoted by $y^* = \left\{y_1^*, \ldots, y_{267}^*\right\}$ and the collection of all unknown subdive behaviours is denoted by $x^* = \left\{x_1^*, \ldots, x_{267}^*\right\}$.

3.2. Model Definition and Selection

The primary goal of this case study is to jointly estimate the dive types and subdive behaviours of this killer whale, so we only consider models with both a coarse-scale and a fine-scale component to describe the kinematic data. Defining a suitable hierarchical model involves selecting an appropriate number of hidden states, model structure, and set of emission distributions for both the coarse- and fine-scale observations.

We do not use information criteria to select the number of dive types N since these metrics tend to overestimate the number of behavioural states in biological processes (Pohle et al., 2017). We instead plot the duration of each dive versus the duration of the dive preceding it (y_t vs. y_{t-1} for $t \geq 2$). This type of visualization is known as a lag plot. If the emission distributions of the hidden states are well separated, a lag plot should reveal N distinct patterns, where each pattern corresponds to one dive type (Lawler et al., 2019). This is unfortunately not the case for our killer whale data, as there is one cluster of data centred at approximately $y_t = y_{t-1} = 30$ s. However, longer dives appear to be characterized by bouts of less "wiggly" behaviour in the acceleration data compared to shorter dives, so we choose $N = 2$ to differentiate these dive types. The absence of a more principled method to select N highlights the importance of model validation techniques in lieu of information criteria (see Section 3.4). Lag plots reveal no significant autocorrelation between dive duration observations (see Figure S1 in the Supplementary Material A), and visual inspection shows no obvious complicated dependence. Therefore, we select a simple HMM to model the coarse-scale process since neither a CarHMM nor a moving-window transformation is called for. Given that dive t is of type i, we assume that the dive duration Y_t follows a gamma distribution with unknown parameters $\mu^{(i)}$ and $\sigma^{(i)}$: $\mathbb{E}(Y_t \mid X_t = i) = \mu^{(i)}$ and $\mathbb{V}(Y_t \mid X_t = i) = \left(\sigma^{(i)}\right)^2$. This is consistent with previous studies, including Leos-Barajas et al. (2017).

We then select a model corresponding to the fine-scale observations of acceleration. Similar to the coarse model, we rely on lag plots and visual inspection to select $N^* = 3$ subdive states. Although N^* is selected heuristically, we test the validity of this model in Section 3.4. In contrast to the coarse-scale observations, the fine-scale acceleration data exhibit significant sinusoidal behaviour. Thus, we transform each fine-scale observation sequence y_t^* into $\tilde{y}_t^*$ using Equation (7) with a window size of $h = 100$ (2 s) and a maximum frequency of $\tilde{\omega} = 10$ (5 Hz). We then have that $\tilde{y}_{t,\tilde{t}^*}^* = \left\{\tilde{a}_{t,\tilde{t}^*}^*, \tilde{w}_{t,\tilde{t}^*}^*\right\}$, where $\tilde{a}_{t,\tilde{t}^*}^*$ is a three-dimensional vector of component-wise average acceleration and $\tilde{w}_{t,\tilde{t}^*}^*$ is a scalar describing the "wiggliness" of a particular window. Even after transforming the raw acceleration data, there is still strong autocorrelation within each component of $\tilde{a}_{t,\tilde{t}^*}^*$ (see Figure S1 in the Supplementary Material A). Therefore, we choose a CarHMM as the fine-scale model.

We then select the specific emission distribution of $\tilde{Y}_{t,\tilde{t}^*}^*$ for all dive types and subdive states. First, we assume that $\tilde{W}_{t,\tilde{t}^*}^*$ and all three components of $\tilde{A}_{t,\tilde{t}^*}^*$ are independent of one another when conditioned on the dive types and subdive states. To reduce model complexity, we also assume that the three sets of fine-scale emission parameters are shared across the two dive types $\left(\text{i.e., } \theta^{*(1,i^*)} = \theta^{*(2,i^*)} = \theta^{*(\cdot,i^*)} \text{ for } i^* = 1, 2, 3\right)$. This implies that the subdive states within dive type 1 have the same interpretation as those within dive type 2. To specify the emission distribution of $\tilde{A}_{t,\tilde{t}^*}^*$, consider the sequence $\left\{\tilde{A}_{t,1}^*, \ldots, \tilde{A}_{t,T_t^*}^*\right\}$ for a particular dive t. We assume

that each of the three components of this sequence is normally distributed as in Equations (3) and (4), and we assume that all components are independent of one another when conditioned on the subdive states $\tilde{X}_t^* = \left\{\tilde{X}_{t,1}^*, \ldots, \tilde{X}_{t,\tilde{T}_t^*}^*\right\}$. Each component is assumed to have its own mean and variance parameters, but all components share the same autocorrelation parameter. Thus, the distribution of $\tilde{A}_{t,\tilde{t}^*}^*$ given $X_{t,\tilde{t}^*}^* = i^*$ has the parameters $\boldsymbol{\mu}_A^{*(\cdot,i^*)} \in \mathbb{R}^3$, $\boldsymbol{\sigma}_A^{*(\cdot,i^*)} \in \mathbb{R}^3$, and $\boldsymbol{\phi}_A^{*(\cdot,i^*)} \in [0,1]$. To specify the emission distribution of $\tilde{W}_{t,\tilde{t}^*}^*$, we assume that, given $\tilde{X}_{t,\tilde{t}^*}^* = i^*$, $\tilde{W}_{t,\tilde{t}^*}^*$ follows a gamma distribution parameterized by its mean $\mu_W^{*(\cdot,i^*)}$ and standard deviation $\sigma_W^{*(\cdot,i^*)}$. In addition, $\tilde{W}_{t,1}^*, \ldots, \tilde{W}_{t,T_t^*}^*$ are assumed to be independent of one another given the subdive state sequence $\left\{\tilde{X}_{t,1}, \ldots, \tilde{X}_{t,T_t^*}^*\right\}$. We do not include $\tilde{W}_{t,\tilde{t}^*-1}^*$ in the distribution of $\tilde{W}_{t,\tilde{t}^*}^*$ because the autocorrelation evident from the lag plot is not severe and may be explained by subsequent observations occurring within the same subdive state.

In total, the parameters to be estimated are the transition probability matrices Γ and $\Gamma^* = \left\{\Gamma^{*(1)}, \Gamma^{*(2)}\right\}$, the coarse-scale emission parameters $\theta = \left\{\mu^{(1)}, \sigma^{(1)}, \mu^{(2)}, \sigma^{(2)}\right\}$, and the fine-scale emission parameters $\theta^* = \left\{\theta^{*(\cdot,1)}, \theta^{*(\cdot,2)}, \theta^{*(\cdot,3)}\right\}$, where $\theta^{*(\cdot,i^*)} = \left\{\mu_A^{*(\cdot,i^*)}, \sigma_A^{*(\cdot,i^*)}, \phi_A^{*(\cdot,i^*)}, \mu_W^{*(\cdot,i^*)}, \sigma_W^{*(\cdot,i^*)}\right\}$. Recall that $\theta^{*(\cdot,i^*)}$ is the set of parameters describing the distribution of $\tilde{Y}_{t,\tilde{t}^*}^*$ conditioned on $\tilde{X}_{t,\tilde{t}^*}^* = i^*$. We refer to this final model as the CarHHMM-DFT since it includes a CarHMM, HHMM, and DFT-based transformation. The likelihood of this model is easily calculated using the forward algorithm and can be maximized with respect to the parameters above (see Appendix for details). Figure 3 shows the dependence structure of the full CarHHMM-DFT.

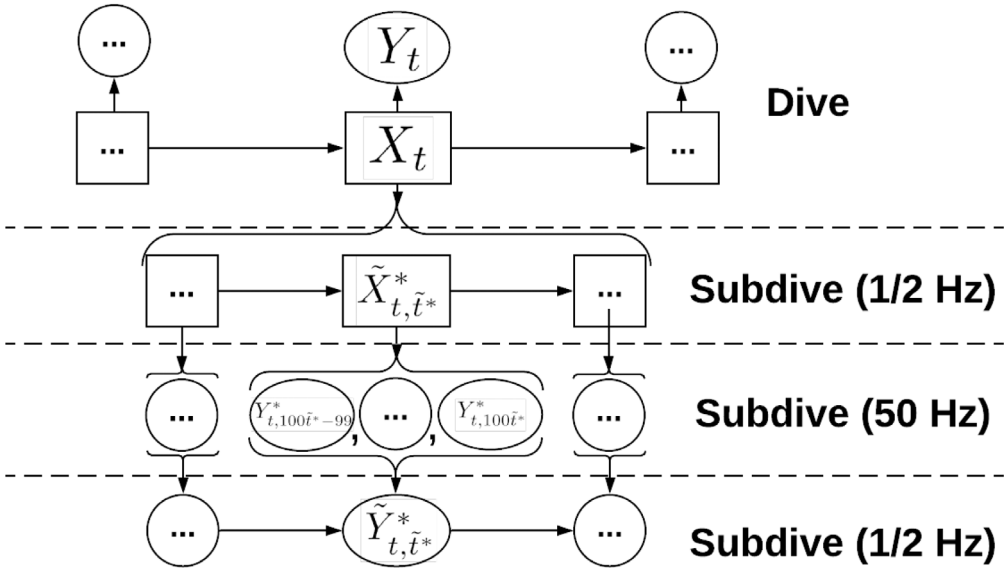

FIGURE 3: Graphical representation of the CarHHMM-DFT used in the simulation and case studies. The type of dive t is denoted by X_t, and Y_t represents the associated dive duration. The raw acceleration vector associated with dive t and time stamp t^* is denoted by Y_{t,t^*}^*. The subdive state of the killer whale during dive t and window $\tilde{t}^*$ is denoted by $\tilde{X}_{t,\tilde{t}^*}^*$, and the corresponding transformed observation is denoted by $\tilde{Y}_{t,\tilde{t}^*}^*$.

In addition to the CarHHMM-DFT, we consider three variations for comparison. As in the full model, each of the following models assumes that all components of $\tilde{Y}^*_{t,\tilde{l}^*}$ are conditionally independent of one another given the dive types and subdive states:

1. An HHMM-DFT, which models the coarse-scale observations with an HMM and transforms the fine-scale observations using Equation (7) but models $\tilde{Y}^*_{t,\tilde{l}^*}$ as emissions from a simple HMM rather than a CarHMM;
2. A CarHHMM, which models the coarse-scale observations with an HMM, transforms the fine-scale observations using Equation (7), and models $\tilde{A}^*_{t,\tilde{l}^*}$ as emissions of a CarHMM;
3. A CarHMM-DFT, which models the coarse-scale observations as an independent and identically distributed sequence of dives, transforms the fine-scale observations using Equation (7), and models $\tilde{Y}^*_{t,\tilde{l}^*}$ as emissions of a CarHMM.

Each of the three candidate models above leaves out one important aspect of the full CarHHMM-DFT: the HHMM-DFT assumes there is no autocorrelation between fine-scale observations; the CarHHMM does not incorporate "wiggliness" $\left(\tilde{W}^*_{t,\tilde{l}^*}\right)$; and the CarHMM-DFT lacks a hierarchical structure and assumes that there is only one dive type.

We estimate the parameters of all four models using the data shown in Figure 2 and direct likelihood maximization using the SciPy package in Python (Virtanen et al., 2019). Each model is fit 100 times using random initializations, and we keep the parameter estimates corresponding to the maximum likelihood of each optimization routine (Zucchini, Macdonald & Langrock, 2016). Likelihood maximization is performed on the Cedar Compute Canada cluster with one CPU and 8 GB of dedicated memory. Most but not all initializations converge to the same parameter estimates, highlighting the need to perform multiple optimizations with a variety of initial guesses. Plots of the likelihood surface near the maximum likelihood estimates are shown in Section 2 of Supplementary Material A.

3.3. Case Study Results

We first report results from the full CarHHMM-DFT in detail and assess the quality of the fit. We then compare these results with those from the other candidate models.

The coarse-scale parameter estimates suggest that the killer whale has at least two distinct dive behaviours (see Table 1 and Figure 4). Dive type 1 corresponds to shorter and shallower dives, which likely reflect resting, travelling, and, to a lesser extent, searching for prey. Dive type 2 is longer and deeper and may be associated with behaviours such as hunting (Tennessen et al., 2019a), but it is unclear whether any of the dives in this study are successful foraging dives. No dive in this data set has a maximum depth greater than 30 m, while Wright et al. (2017), in a study of killer whales in Johnstone Strait, found that most prey captures occur at depths deeper than 100 m. However, the killer whale studied here was tagged north of Johnstone Strait in Queen Charlotte Sound, and unpublished data collected by the authors suggest that prey-capture events can occur near the surface. Dive type 2 could also be associated with behaviours such as socializing, which can take place several meters below the surface (Tennessen et al., 2019a).

The means of "wiggliness" $\left(\tilde{W}^*_{t,\tilde{l}^*}\right)$ associated with each subdive state are separated by an order of magnitude (see Table 1 and Figure 4). Subdive state 1 has the smallest mean corresponding to $\tilde{W}^*_{t,\tilde{l}^*}$ and the smallest standard deviation corresponding to $\tilde{A}^*_{t,\tilde{l}^*}$. It also has the highest autocorrelation in $\tilde{A}^*_{t,\tilde{l}^*}$. This implies less overall activity and more consistent acceleration compared to the other subdive states. Subdive state 2 has a mean "wiggliness" one order of magnitude higher than that of subdive state 1 and its acceleration has about twice the standard deviation compared to subdive state 1. The autocorrelation of acceleration is also slightly lower than in subdive state 1. We therefore hypothesize that subdive state 2 corresponds to fluking

TABLE 1: Estimates and standard errors for the parameters of the distribution of dive duration (Y_t), acceleration $\left(\tilde{A}^*_{t,\tilde{t}^*}\right)$, and "wiggliness" $\left(\tilde{W}^*_{t,\tilde{t}^*}\right)$ of the killer whale kinematic data using the full CarHHMM-DFT.

Feature	Dive type/ subdive state	Parameter estimate		
		$\hat{\mu}$	$\hat{\sigma}$	$\hat{\phi}$
Dive duration (s)—Y_t	1	27.342 ± 0.633	10.961 ± 0.560	—
	2	127.548 ± 11.341	63.888 ± 9.032	—
x-Acc. (m/s^2)—$\left(\tilde{A}^*_{t,\tilde{t}^*}\right)_x$	1	0.449 ± 0.030	0.039 ± 0.001	0.968 ± 0.002
	2	0.210 ± 0.012	0.096 ± 0.002	0.829 ± 0.007
	3	0.232 ± 0.035	0.296 ± 0.010	0.607 ± 0.023
y-Acc. (m/s^2)—$\left(\tilde{A}^*_{t,\tilde{t}^*}\right)_y$	1	0.450 ± 0.038	0.051 ± 0.001	0.968 ± 0.002
	2	0.437 ± 0.012	0.094 ± 0.002	0.829 ± 0.007
	3	0.366 ± 0.042	0.365 ± 0.012	0.607 ± 0.023
z-Acc. (m/s^2)—$\left(\tilde{A}^*_{t,\tilde{t}^*}\right)_z$	1	-0.691 ± 0.043	0.058 ± 0.001	0.968 ± 0.002
	2	-0.573 ± 0.014	0.111 ± 0.002	0.829 ± 0.007
	3	-0.303 ± 0.041	0.354 ± 0.012	0.607 ± 0.023
Wiggliness—$\tilde{W}^*_{t,\tilde{t}^*}$	1	34.015 ± 0.368	22.986 ± 0.378	—
	2	490.068 ± 5.584	502.558 ± 6.776	—
	3	9154.156 ± 220.765	$13{,}538.747 \pm 354.281$	—

Note: Figures following $\pm$ refer to standard errors estimated using the observed information matrix.

(active swimming), as strong sinusoidal behaviour in acceleration is characteristic of fluking in marine mammals (Simon, Johnson & Madsen, 2012). Finally, the mean of $\tilde{W}^*_{t,\tilde{t}^*}$ and variance of $\tilde{A}^*_{t,\tilde{t}^*}$ in subdive state 3 are both much higher than in the other two states, and the autocorrelation of $\tilde{A}^*_{t,\tilde{t}^*}$ is also much lower. This corresponds to vigorous swimming activity, especially as the killer whale ends a dive (see Figure 5).

The estimated transition probability matrix and associated stationary distribution on the coarse scale (i.e., between dives) are

$$\hat{\Gamma} = \begin{pmatrix} 0.847 & 0.153 \\ 0.914 & 0.086 \end{pmatrix}$$

and $\hat{\delta} = (0.857, 0.143)$. The estimated transition probability matrices and stationary distributions on the fine scale are

$$\hat{\Gamma}^{*(1)} = \begin{pmatrix} 0.745 & 0.253 & 0.002 \\ 0.080 & 0.869 & 0.052 \\ 0.000 & 0.229 & 0.771 \end{pmatrix}, \qquad \hat{\Gamma}^{*(2)} = \begin{pmatrix} 0.886 & 0.114 & 0.000 \\ 0.150 & 0.815 & 0.036 \\ 0.000 & 0.225 & 0.775 \end{pmatrix},$$

$\hat{\delta}^{*(1)} = (0.202, 0.649, 0.149)$, and $\hat{\delta}^{*(2)} = (0.531, 0.405, 0.064)$ for dive types 1 and 2. In summary, about 86% of dives are short dives of type 1. The whale performs an average of 6.54 short

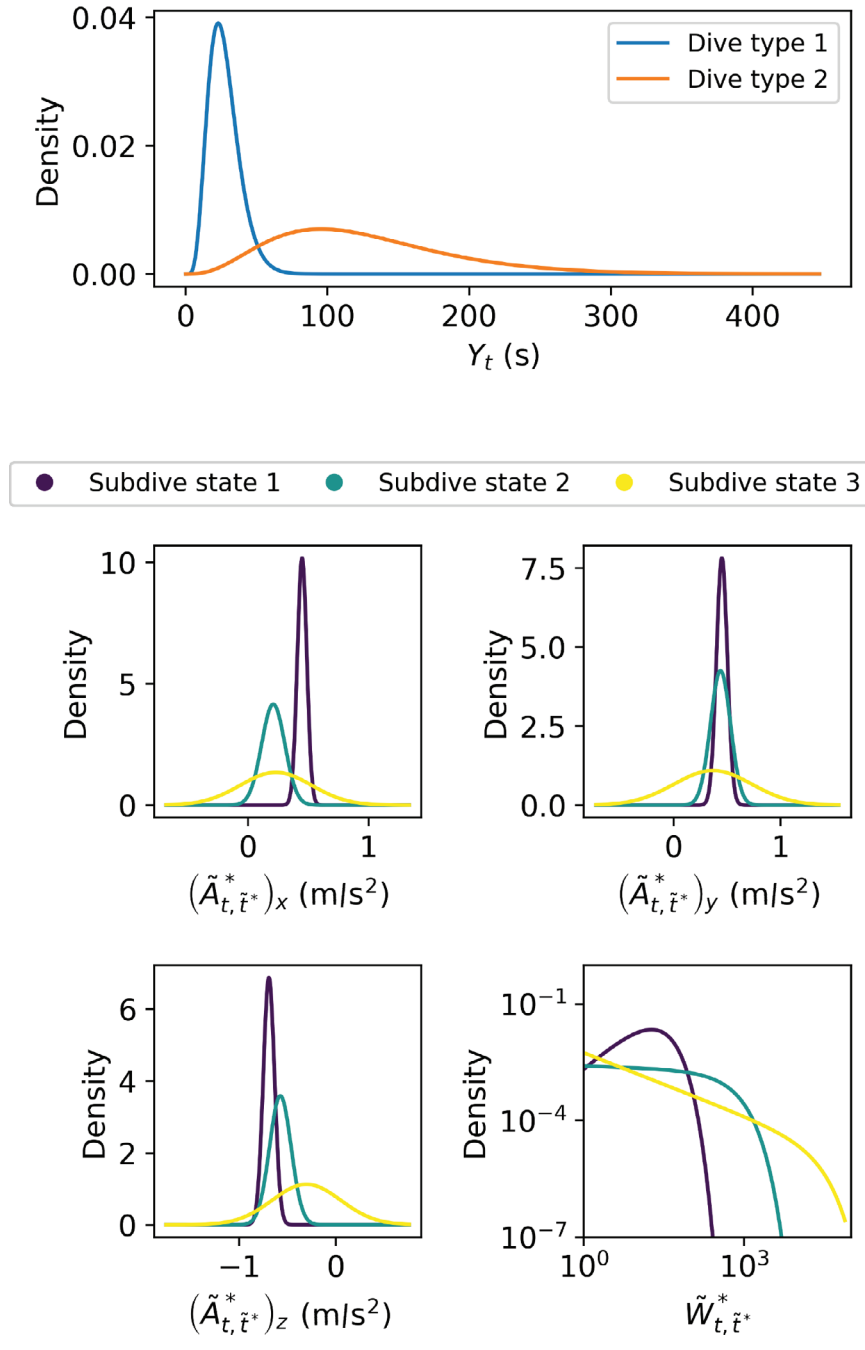

FIGURE 4: Estimated gamma densities of killer whale dive duration (Y_t) (top), estimated normal conditional densities of killer whale acceleration $\left(\tilde{A}^*_{t,\tilde{t}^*} \mid \tilde{A}^*_{t,\tilde{t}^*-1} = \mu_A^{*(\cdot,\tilde{t}^*)}\right)$ (middle and bottom left), and estimated gamma densities of "wiggliness" $\left(\tilde{W}^*_{t,\tilde{t}^*}\right)$ plotted on a log−log scale (bottom right). The densities of Y_t correspond to dive types 1 and 2, while the densities of $\tilde{A}^*_{t,\tilde{t}^*}$ and $\tilde{W}^*_{t,\tilde{t}^*}$ correspond to subdive states 1, 2, and 3. The densities are estimated by fitting the CarHHMM-DFT to the case study data (see Table 1).

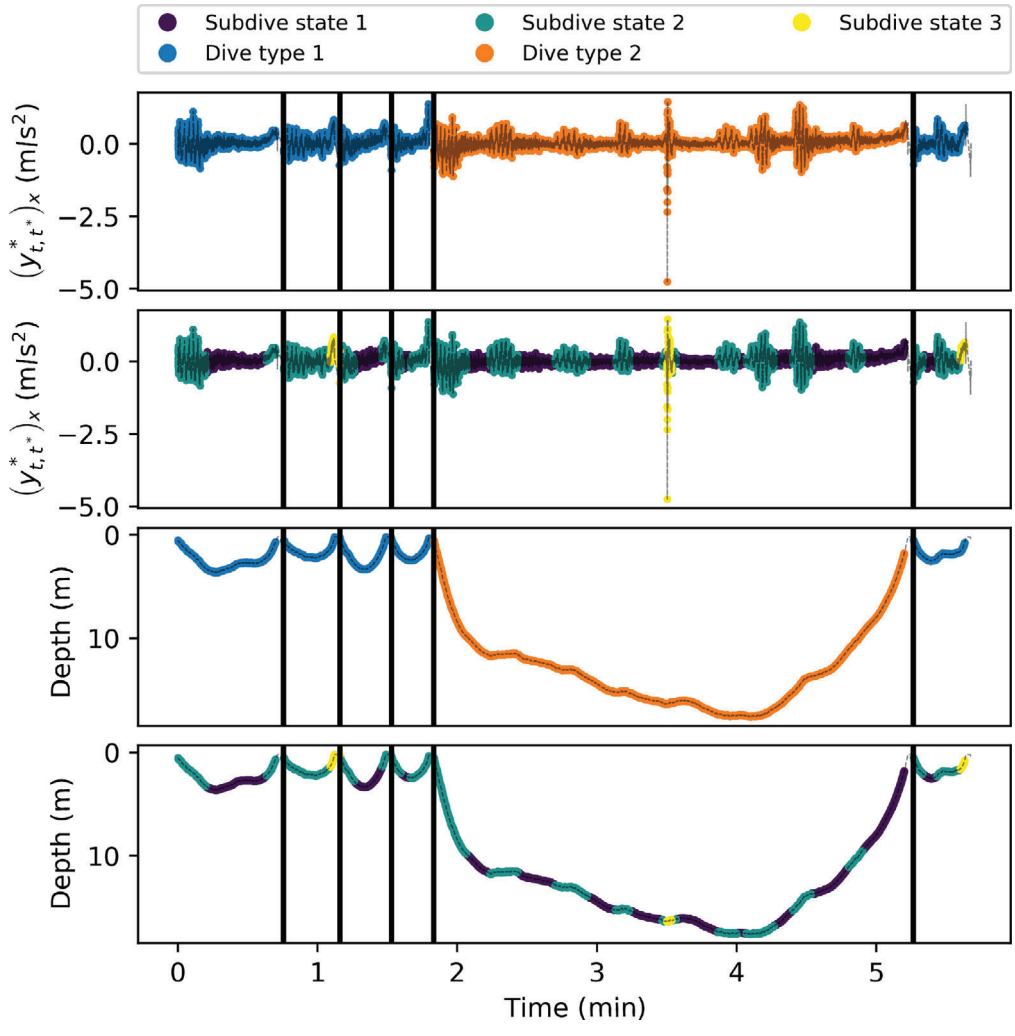

FIGURE 5: The x-component of acceleration, $(y^*_{t,t^*})_x$ (top two panels), and dive depth (bottom two panels) of a northern resident killer whale for a sequence of six selected dives. Each panel is partitioned into dives by vertical black lines. Curve colour in the first and third panels corresponds to estimated dive type, while curve colour in the second and fourth panels corresponds to the estimated subdive state. Both dive type and subdive state are estimated by fitting the CarHHMM-DFT to the data and applying the forward–backward algorithm to determine the hidden state with the highest probability.

type 1 dives before switching to dive type 2 and an average of 1.09 longer type 2 dives before switching back to dive type 1. This finding is consistent with those of Williams & Noren (2009) and Tennessen et al. (2019a), both of which describe common bouts of short resting dives before a longer, more energy-intensive deep dive. Furthermore, this killer whale is in the less active subdive state 1 only 20% of the time during a dive of type 1, compared to 53% of the time during a dive of type 2. Less active swimming behaviour is consistent with the need for marine mammals to conserve energy when diving to greater depths and holding their breath for long periods (Williams, Haun & Friedl, 1999; Hastie, Rosen & Trites, 2006). Figure 5 shows the

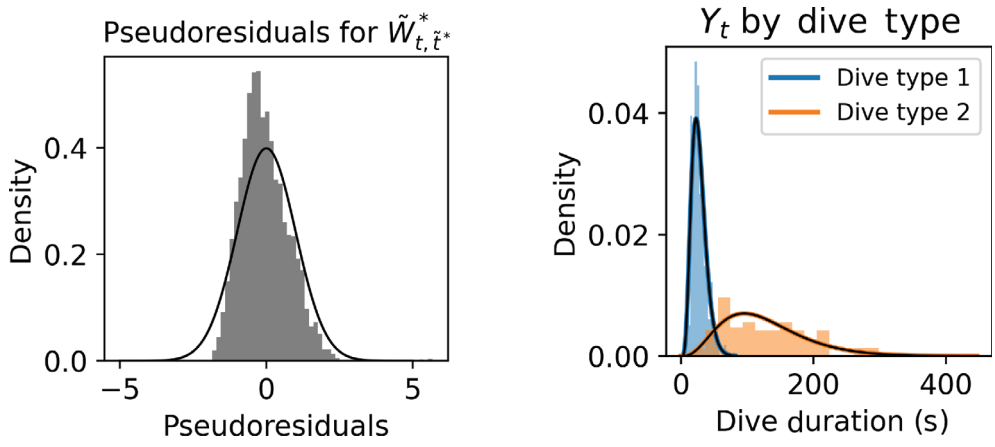

FIGURE 6: Pseudoresiduals of "wiggliness" plotted over a standard normal density (left) and weighted empirical distributions of dive duration Y_t plotted over the corresponding fitted gamma distributions (right). Both plots are generated by fitting the CarHHMM-DFT to the killer whale case study data and applying the forward–backward algorithm.

decoded dive behaviour of six selected dives. Section 4 of Supplementary Material A also shows the probability of each dive type and subdive state given the data and the fitted model.

3.4. Model Validation

We use two visual tools to evaluate the CarHHMM-DFT: pseudoresidual plots and empirical histograms. The pseudoresidual of a coarse-scale observation y_t is $\Phi^{-1}\left(\Pr\left(Y_t < y_t \mid \left\{Y_1, \ldots, Y_T, \tilde{Y}^*_1, \ldots, \tilde{Y}^*_T\right\} \setminus \{Y_t\}\right)\right)$, and the pseudoresidual of a fine-scale observation $\tilde{y}^*_{t,\tilde{t}^*}$ is $\Phi^{-1}\left(\Pr\left(\tilde{Y}^*_{t,\tilde{t}^*} < \tilde{y}^*_{t,\tilde{t}^*} \mid \left\{Y_1, \ldots, Y_T, \tilde{Y}^*_1, \ldots, \tilde{Y}^*_T\right\} \setminus \left\{Y^*_{t,\tilde{t}}\right\}\right)\right)$, where Φ is the cumulative distribution function of the standard normal distribution. If the model is correct, then all pseudoresiduals are independent and follow the standard normal distribution. Histograms of the pseudoresiduals mostly support that the CarHHMM-DFT is well specified. One exception is $\tilde{W}^*_{t,\tilde{t}^*}$, whose pseudoresiduals are noticeably right-skewed (see Figure 6). This implies that the true distribution of $\tilde{W}^*_{t,\tilde{t}^*}$ may be heavier tailed than the gamma distribution used in the case study. See Sections 5–7 of Supplementary Material A for pseudoresidual plots corresponding to all observations and models.

We also plot histograms of dive duration corresponding to each dive type in Figure 6. Each observation of dive duration is weighted by the estimated probability that it corresponds to a particular dive type as decoded by the forward–backward algorithm. This procedure results in two histograms—one corresponding to dive type 1, and the other corresponding to dive type 2. Each histogram is plotted together with the corresponding emission distribution estimated by the CarHHMM-DFT. Analogous histograms corresponding to the fine-scale observations are contained in Sections 6 and 7 of Supplementary Material A. Our results mostly show that the CarHHMM-DFT explains the data well, but there are some exceptions. In particular, histograms corresponding to subdive state 3 show that $\tilde{A}^*_{t,\tilde{t}^*}$ has heavier tails compared to a normal distribution. This indicates the existence of rare events corresponding to exceptionally sudden changes in the acceleration of the killer whale. These outliers are potential subjects for future study and may indicate biologically relevant phenomena such as prey capture (Tennessen et al., 2019b).

3.5. Comparison with Candidate Models

The HHMM-DFT, which ignores autocorrelation in acceleration, decodes dive types and subdive states similarly to the CarHHMM-DFT, but is less likely to categorize any given behaviour as subdive state 3 (see Figures S11 and S12 in the Supplementary Material A). In addition, for all three components of $\sigma_A^{*(\cdot,1)}$, $\sigma_A^{*(\cdot,2)}$, and $\sigma_A^{*(\cdot,3)}$, the HHMM-DFT produces estimates which are significantly larger than those of the CarHHMM-DFT. The estimated uncertainties of the three components of $\hat{\mu}_A^{*(\cdot,1)}$, $\hat{\mu}_A^{*(\cdot,2)}$, and $\hat{\mu}_A^{*(\cdot,3)}$ are also less than half of those for the CarHHMM-DFT (see Tables S1 and S2 in the Supplementary Material A). This suggests that the HHMM-DFT overlooks some autocorrelation in the data, and that the HHMM-DFT may be overconfident in its parameter estimates compared to the full CarHHMM-DFT.

The CarHHMM does not model the "wiggliness" of the acceleration data, so it regularly fails to pick up obvious behavioural changes corresponding to the periodicity shown in Figure 5 (see Figure S13 in the Supplementary Material A). These results essentially disqualify the CarHHMM as a viable model for this data set. The acceleration pseudoresiduals are also light-tailed relative to a normal distribution (see Figure S23 in the Supplementary Material A).

Finally, the CarHMM-DFT, which lacks a hierarchical structure, produces fine-scale parameter estimates and subdive state estimates similar to those of the CarHHMM-DFT. However, the former's lack of hierarchical structure means that it fails to differentiate between short and long dives. This model therefore does not infer the dive-level Markov chain or the relationship between the dive types and subdive states. For example, the CarHMM-DFT does not indicate that the whale is more likely to be in subdive state 1 when engaged in longer dives compared to shorter dives.

For a more complete set of results for each of the candidate models, see Supplementary Material A.

4. SIMULATION STUDY

We perform a simulation study based on data generated from the full CarHHMM-DFT, as defined in Section 3.2, to evaluate each candidate model when the ground truth is known. The parameters used to generate the data are based on those estimated in the case study (see Table 1), with slight modifications made for simplicity. In particular, we set $\tilde{A}^*_{t,\tilde{t}^*}$ to a scalar instead of a three-dimensional vector. We then fit all four models to the simulated data. Metrics used to evaluate each model include hidden state decoding accuracy, bias in parameter estimates, empirical standard errors of parameter estimates, and fitting times. To assess the accuracy of uncertainty estimates, we also compare the empirical standard errors of a given model's parameter estimates with the standard errors estimated using the inverse of the observed Fisher information.

4.1. Simulation Procedure

We generate 500 independent training data sets using the CarHHMM-DFT as a generative model. Each training data set consists of a sequence of 100 curves, which we call a sequence of killer whale dives. Each dive can be one of $N = 2$ dive types based on a Markov chain with transition probability matrix

$$\Gamma = \begin{pmatrix} 0.847 & 0.153 \\ 0.914 & 0.086 \end{pmatrix}.$$

Dive duration is gamma-distributed, and the coarse-scale emission parameters are $\mu^{(1)} = 27.34$ s, $\sigma^{(1)} = 10.96$ s, $\mu^{(2)} = 127.55$ s, and $\sigma^{(2)} = 63.89$ s. After generating the dive durations for all 100 dives in a data set, dive t is broken into a sequence of $\tilde{T}^*_t = \lfloor Y_t/2 \rfloor$ 2-s windows, where

 DOI: 10.1002/cjs.11673

the last $Y_t - 2\tilde{T}_t^*$ seconds of each simulated dive are ignored. Each 2-s segment is assigned one of $N^* = 3$ behaviours according to a fine-scale Markov chain $\tilde{X}_t^* = \left\{\tilde{X}_{t,1}^*, \ldots, \tilde{X}_{t,\tilde{T}_t^*}^*\right\}$ with the transition probability matrices

$$\Gamma^{*(1)} = \begin{pmatrix} 0.745 & 0.253 & 0.002 \\ 0.080 & 0.868 & 0.052 \\ 0.000 & 0.229 & 0.771 \end{pmatrix} \quad \text{and} \quad \Gamma^{*(2)} = \begin{pmatrix} 0.886 & 0.139 & 0.000 \\ 0.150 & 0.815 & 0.035 \\ 0.000 & 0.225 & 0.775 \end{pmatrix}$$

for dive types 1 and 2, respectively. Instead of generating the raw observations Y_{t,t^*}^*, we directly simulate the fine-scale transformed observations $\tilde{Y}_{t,\tilde{t}^*}^* = \left\{\tilde{A}_{t,\tilde{t}^*}^*, \tilde{W}_{t,\tilde{t}^*}^*\right\}$. Recall from Section 3.2 that we must specify the mean, standard deviation, and autocorrelation parameters corresponding to $\left\{\tilde{A}_{t,1}^*, \ldots, \tilde{A}_{t,\tilde{T}_t^*}^*\right\}$ as well as the mean and standard deviation parameters corresponding to $\left\{\tilde{W}_{t,1}^*, \ldots, \tilde{W}_{t,\tilde{T}_t^*}^*\right\}$. We select the following parameters in line with the results from the case study:

1. $\mu_A^{*(\cdot,1)} = 0.0$ s, $\sigma_A^{*(\cdot,1)} = 0.05$ s, $\phi_A^{*(\cdot,1)} = 0.97$, $\mu_W^{*(\cdot,1)} = 34.01$, and $\sigma_W^{*(\cdot,1)} = 22.99$.
2. $\mu_A^{*(\cdot,2)} = 0.1$ s, $\sigma_A^{*(\cdot,2)} = 0.1$ s, $\phi_A^{*(\cdot,2)} = 0.83$, $\mu_W^{*(\cdot,2)} = 490.06$, and $\sigma_W^{*(\cdot,2)} = 502.56$.
3. $\mu_A^{*(\cdot,3)} = 0.2$ s, $\sigma_A^{*(\cdot,3)} = 0.3$ s, $\phi_A^{*(\cdot,3)} = 0.61$, $\mu_W^{*(\cdot,3)} = 9154.16$, and $\sigma_W^{*(\cdot,3)} = 13{,}538.75$.

It is not possible to uniquely reconstruct the raw accelerometer data Y^* from $\tilde{Y}^*$ alone, but we describe one possible mapping from $\tilde{Y}^*$ to Y^* in the Appendix. Figure S1 in the Supplementary Material B shows one realization of $\tilde{Y}^*$ for five dives of one simulated data set along with the corresponding reconstructed realizations of Y^*.

The two simulated dive types differ in that dives of type 1 are much shorter on average (27 s) than dives of type 2 (128 s). The three simulated subdive states differ primarily because μ_W^* and σ_W^* are much higher for subdive state 3 than for subdive state 2, which in turn are much higher than for subdive state 1. These larger parameter values correspond to much more vigorous and variable periodic behaviour in the acceleration data.

We calculate the maximum likelihood estimates $\left\{\hat{\theta}, \hat{\Gamma}, \hat{\theta}^*, \hat{\Gamma}^*\right\}$ for all four candidate models for each of the 500 data sets using an optimization procedure similar to that in the case study. For each of the 500 training data sets, we simulate a test data set to assess how well each model predicts the hidden states as follows: Each test data set consists of a sequence of 100 dives and is created from the generative model with the true parameters $\{\theta, \Gamma, \theta^*, \Gamma^*\}$. To assess the coarse-scale hidden state prediction, we estimate $p_t(i \mid y, \tilde{y}^*) = \Pr(X_t = i \mid Y = y, \tilde{Y}^* = \tilde{y}^*)$ for $i = 1, 2$ and $t = 1, \ldots, 100$ using the test set observations $(y, \tilde{y}^*)$ and the training set maximum likelihood estimates. These estimates are found using the forward–backward algorithm (Zucchini, Macdonald & Langrock, 2016). We compare these estimated conditional probabilities to $\left\{x_1, \ldots, x_{100}\right\}$, the true coarse-scale state realizations in the test data, by calculating the average dive decoding accuracy for a single training/test data set pair, $\sum_{t=1}^{100} \hat{p}_t(x_t \mid y, \tilde{y}^*)/100$. We then report the average of this measure over the 500 training/test data set pairs. Analogously, to assess fine-scale state prediction, we estimate $p_{t,\tilde{t}^*}^*(i^* \mid y, \tilde{y}^*) = \Pr\left(\tilde{X}_{t,\tilde{t}^*}^* = i \mid Y = y, \tilde{Y}^* = \tilde{y}^*\right)$ for $i^* = 1, 2, 3$, $\tilde{t}^* = 1, \ldots, \tilde{T}_t^*$, and $t = 1, \ldots, 100$ using the test set observations, the training set maximum likelihood estimates, and the forward–backward algorithm. Denoting the true fine-scale state realizations from the test data set by $\left\{\tilde{x}_{t,1}^*, \ldots, \tilde{x}_{t,\tilde{T}_t^*}^*\right\}$, we define the overall average subdive decoding accuracy as the average value of $\hat{p}_{t,\tilde{t}^*}^*\left(\tilde{x}_{t,\tilde{t}^*}^* \mid y, \tilde{y}^*\right)$ across all simulated test data sets, dives, and windows. The conditional probabilities are estimated according to each of the four models under study using the maximum likelihood estimates from the training data set in conjunction with the forward–backward algorithm.

TABLE 2: Average decoding accuracies and training times for all models in the simulation study.

Model	Training time (min)	Dive type	Subdive type	Dive accuracy	Subdive accuracy
CarHHMM-DFT	156 ± 67	All	All	0.956 ± 0.028	0.911 ± 0.006
		1	1		0.846 ± 0.027
		1	2	0.973 ± 0.032	0.918 ± 0.011
		1	3		0.860 ± 0.030
		2	1		0.949 ± 0.010
		2	2	0.856 ± 0.097	0.915 ± 0.015
		2	3		0.871 ± 0.057
HHMM-DFT	162 ± 64	All	All	0.959 ± 0.025	0.844 ± 0.024
		1	1		0.761 ± 0.063
		1	2	0.977 ± 0.028	0.845 ± 0.029
		1	3		0.858 ± 0.034
		2	1		0.883 ± 0.066
		2	2	0.851 ± 0.097	0.829 ± 0.041
		2	3		0.876 ± 0.061
CarHHMM	258 ± 106	All	All	0.924 ± 0.127	0.845 ± 0.018
		1	1		0.689 ± 0.050
		1	2	0.933 ± 0.153	0.900 ± 0.031
		1	3		0.673 ± 0.059
		2	1		0.906 ± 0.022
		2	2	0.864 ± 0.134	0.848 ± 0.033
		2	3		0.687 ± 0.105
CarHMM-DFT	45 ± 22	All	All	—	0.912 ± 0.006
		1	1		0.854 ± 0.026
		1	2	—	0.921 ± 0.011
		1	3		0.860 ± 0.032
		2	1		0.943 ± 0.011
		2	2	—	0.919 ± 0.014
		2	3		0.878 ± 0.053

Note: Each of the four models was fit to 500 training data sets composed of 100 simulated dives and tested on test data sets also composed of 100 simulated dives. Reported values are averages, and figures following ± refer to sample standard deviation across the 500 data sets. "All" denotes overall average decoding accuracy.

4.2. Simulation Results

The full CarHHMM-DFT is the best performing model of the four candidates since it is the generating model. Its average dive decoding accuracy is approximately 0.96 and its average subdive decoding accuracy is approximately 0.91. All parameter estimates of emission

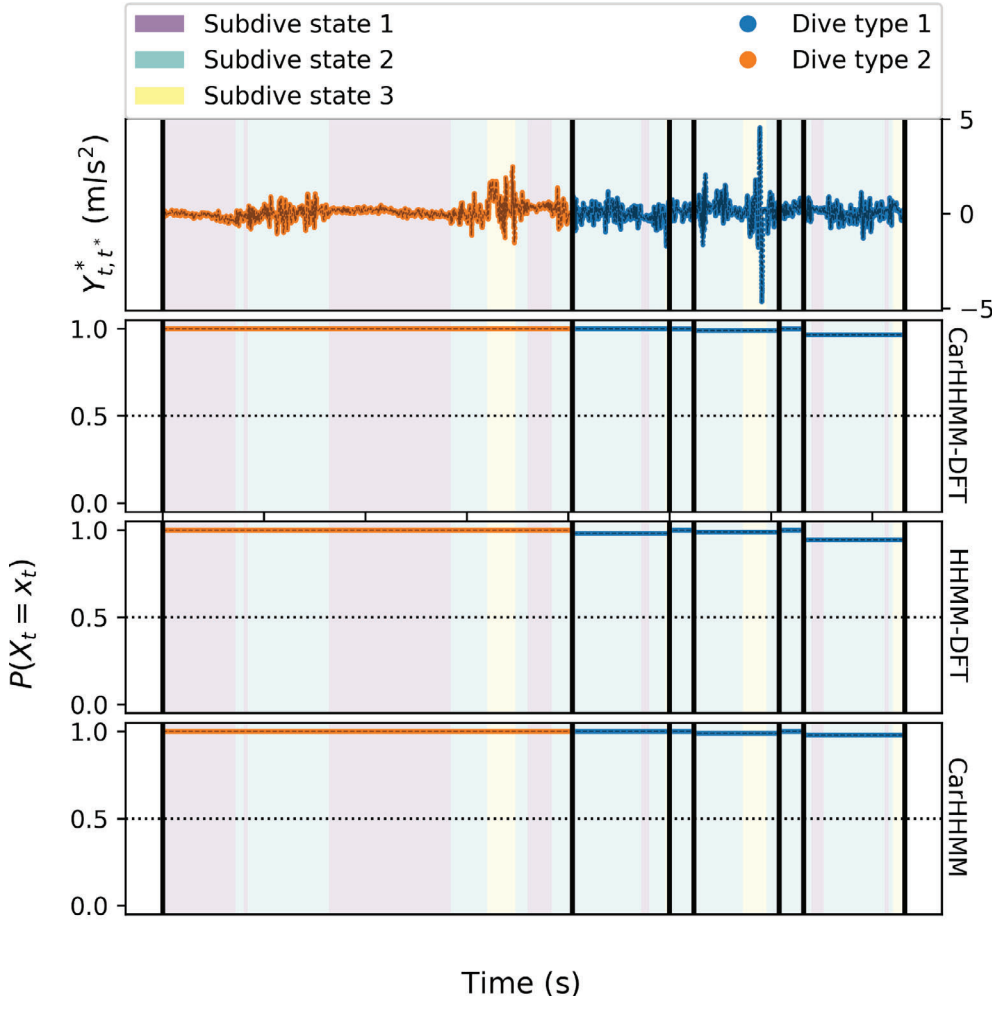

FIGURE 7: Estimated probabilities that each dive t is of type x_t for six selected dives of a simulated data set of killer whale dive behaviour. Each panel is partitioned into dives by vertical black lines. Curve colour corresponds to true dive type, while background colour corresponds to true subdive state. The CarHMM-DFT is omitted because it assumes that there is only one dive type.

distributions (θ and θ^*) and transition probability matrices (Γ and Γ^*) on both the coarse scale and fine scale are either comparable or favourable relative to all other models. The empirical standard errors of all parameter estimates ($\hat{\theta}$, $\hat{\Gamma}$, $\hat{\theta}^*$, and $\hat{\Gamma}^*$) are well approximated by the inverse of the observed Fisher information matrix, although the estimated standard errors tend to be slightly smaller than the empirical standard errors. This underestimation is especially noticeable for parameters associated with the "wiggliness" $\tilde{W}^*_{t,\tilde{t}^*}$, where the empirical standard error can be up to 5 times as large as the estimated standard error. See Tables S2–S6 in the Supplementary Material B for detailed results.

The HHMM-DFT has an average dive decoding accuracy comparable to that of the CarHHMM-DFT (0.96), but its average subdive decoding accuracy is worse by approximately seven percentage points (0.84). The HHMM-DFT's parameter estimates are comparable to

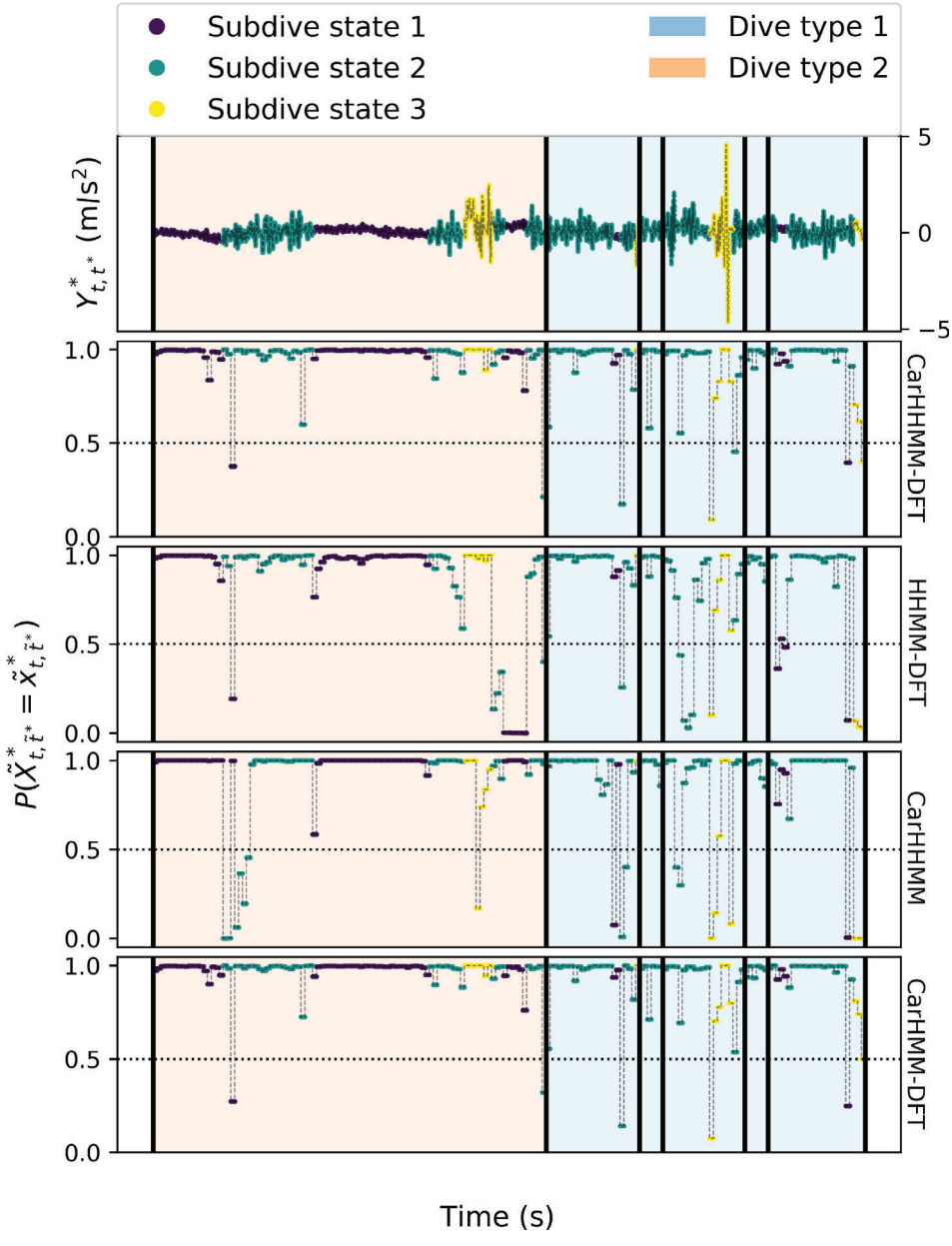

FIGURE 8: Estimated probabilities that each window $(t, \tilde{t}^*)$ corresponds to subdive state $\tilde{x}^*_{t,\tilde{t}^*}$ for six selected dives of a simulated data set of killer whale dive behaviour. Each panel is partitioned into dives by vertical black lines. Curve colour corresponds to true subdive state, while background colour corresponds to true dive type.

those of the CarHHMM-DFT with the notable exception that the former greatly overestimates $\sigma_A^{*(\cdot,1)}$ and $\sigma_A^{*(\cdot,2)}$ and slightly overestimates $\sigma_A^{*(\cdot,3)}$. In addition, the estimated standard errors of $\hat{\mu}_A^{*(\cdot,1)}$, $\hat{\mu}_A^{*(\cdot,2)}$, $\hat{\mu}_A^{*(\cdot,3)}$, $\hat{\sigma}_A^{*(\cdot,1)}$, $\hat{\sigma}_A^{*(\cdot,2)}$, and $\hat{\sigma}_A^{*(\cdot,3)}$ are much smaller than the associated empirical standard errors (see Table S3 in the Supplementary Material B). These results suggest that

estimates of standard deviation can be too large and estimates of standard errors can be too small when autocorrelation is ignored. This finding is consistent with the results of the case study, where the HHMM-DFT produced larger estimates of σ_A^* and smaller estimates of standard error compared to the CarHHMM-DFT. When standard errors are underestimated, the associated confidence intervals are too narrow, and so researchers may be overconfident in their parameter estimates.

The CarHHMM is the worst performing model in terms of its dive decoding accuracy, which is approximately four percentage points worse than that of the CarHHMM-DFT (0.92). The former's average subdive decoding accuracy is approximately six percentage points worse than that of the CarHHMM-DFT (0.85). This result is consistent with our expectations because the CarHHMM does not model the "wiggliness" of the fine-scale process, which is the most distinct difference between the subdive states. In addition to its relatively poor average decoding accuracy, the CarHHMM is also the worst of the four candidate models at estimating emission parameters and transition probability matrices. Estimates associated with subdive states 2 and 3 ($\theta^{*(\cdot,2)}$ and $\theta^{*(\cdot,3)}$) are especially poor. See Supplementary Material B for more-detailed results.

Finally, the CarHMM-DFT is nearly identical to the CarHHMM-DFT in terms of average subdive decoding accuracy, fine-scale parameter biases, and both estimated and empirical standard error for the fine-scale parameter estimates. In addition, the time required to fit the CarHMM-DFT is less than one-third that of the other models (see Table 2). However, this model cannot differentiate between dive types, as it assumes that there is only one. The CarHMM-DFT nonetheless fits a (misspecified) single gamma distribution over the dive duration of all dives. The resulting parameter estimates ($\hat{\mu}_Y$ and $\hat{\sigma}_Y$) are highly correlated (see Figure S5 in the Supplementary Material B).

Figures 7 and 8 display five dives of one simulated data set as well as the decoded dive types and subdive states associated with each model. The CarHHMM-DFT and CarHMM-DFT produce similar estimates of subdive state, the HHMM-DFT is slightly more likely to misclassify subdive state 1, and the CarHHMM is more likely to misclassify subdive state 3. All models classify dive type with high accuracy, with the exception of the CarHMM-DFT, which does not estimate dive type.

5. DISCUSSION

Current functional data analysis literature addresses dependence between curves either with multi-level models (Di et al., 2009; Chen & Müller, 2012), which lack a time component, or with functional time series, which overlook the possibility that curves have several distinct "types" (Kokoszka & Reimherr, 2018). Our work addressed these issues and introduced a flexible framework to model functional time-series data using HMMs. We suggested handling temporal dependence between curves by using either an HMM or a CarHMM to model the curve sequence. We then suggested viewing each individual curve as an HMM emission whose distribution is described by a fine-scale model. Here we used a CarHMM as the fine-scale model, but there are a wide range of possible fine-scale models, including a Poisson process or a continuous-time approach similar to that of Michelot & Blackwell (2019). We also incorporated a moving-window transformation at the fine scale to capture intricate dependence structures. Together, the coarse- and fine-scale models make up a hierarchical structure that can account for simultaneous processes taking place at different time scales. Provided that the construction is not overly complex, a hierarchical model created using our method can be both flexible and easy to fit using maximum likelihood estimation. Our method is not intended to minimize prediction error for functional data, but incorporating HMMs into autoregressive predictive models such as those described in Aue, Norinho & Hörmann (2015) and Gao, Shang & Yang (2019) is a promising and natural direction for future study.

We demonstrated the usefulness of this framework using a biomechanical/ecological example, where we used HMMs to classify the coarse- and fine-scale diving behaviour of a northern resident killer whale in Queen Charlotte Sound, off the coast of British Columbia. Our analysis gave a deeper understanding of the killer whale's tri-axial movement and thus its behaviour and energy expenditure (Gleiss, Wilson & Shepard, 2011; Qasem et al., 2012), both of which are important for understanding the foraging ecology and nutritional status of northern resident killer whales (Noren, 2011). Our model is also applicable to many diving animals such as sharks (Adam et al., 2019), seals (Jeanniard du Dot et al., 2016), and porpoises (Leos-Barajas et al., 2017). In addition, since complicated state-switching processes with temporal dependence are common in settings ranging from speech recognition (Juang & Rabiner, 1991) and neuroscience (Langrock et al., 2013) to oceanography (Bulla et al., 2012) and ecology (Adam et al., 2019), we believe that researchers can adapt our methodology for the analysis of a wide range of time-series data in a variety of fields.

ACKNOWLEDGEMENTS

All killer whale data were collected under University of British Columbia Animal Care Permit no. A19-0053 and Fisheries and Oceans Canada Marine Mammal Scientific License for Whale Research no. XMMS 6 2019. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) as well as of Fisheries and Oceans Canada (DFO). This project was supported in part by a financial contribution from the DFO and NSERC (Whale Science for Tomorrow). This research was enabled in part by support provided by WestGrid (www.westgrid.ca) and Compute Canada (www.computecanada.ca). Marie Auger-Méthé and Nancy Heckman thank the NSERC Discovery program, and Marie Auger-Méthé additionally thanks the Canadian Research Chair program. Evan Sidrow thanks the University of British Columbia for funding provided via the Four-Year Doctoral Fellowship program. We are grateful to Dr. Joe Watson for his constructive suggestions.

REFERENCES

Adam, T., Griffiths, C., Leos Barajas, V., Meese, E., Lowe, C., Blackwell, P., Righton, D., & Langrock, R. (2019). Joint modelling of multi-scale animal movement data using hierarchical hidden Markov models. *Methods in Ecology and Evolution*, 10, 1536–1550.

Aue, A., Norinho, D. D., & Hörmann, S. (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association*, 110, 378–392.

Bebbington, M. S. (2007). Identifying volcanic regimes using hidden Markov models. *Geophysical Journal International*, 171, 921–942.

Borchers, D. L., Zucchini, W., Heide-Jørgensen, M. P., Cañadas, A., & Langrock, R. (2013). Using hidden Markov models to deal with availability bias on line transect surveys. *Biometrics*, 69, 703–713.

Börger, L., Bijleveld, A., Fayet, A., Machovsky-Capuska, G., Patrick, S., Street, G., & Wal, E. (2020). Biologging special feature. *Journal of Animal Ecology*, 89, 6–15.

Brumback, B. A. & Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, 93, 961–976.

Bulla, J., Lagona, F., Maruotti, A., & Picone, M. (2012). A multivariate hidden Markov model for the identification of sea regimes from incomplete skewed and circular time series. *Journal of Agricultural, Biological and Environmental Statistics*, 17, 544–567.

Cade, D. E., Barr, K. R., Calambokidis, J., Friedlaender, A. S., & Goldbogen, J. A. (2018). Determining forward speed from accelerometer jiggle in aquatic environments. *Journal of Experimental Biology*, 221, jeb170449.

Chen, K. & Müller, H.-G. (2012). Modeling repeated functional observations. *Journal of the American Statistical Association*, 107, 1599–1609.

Crainiceanu, C. M., Staicu, A.-M., & Di, C.-Z. (2009). Generalized multilevel functional regression. *Journal of the American Statistical Association*, 104, 1550–1561.

de Souza, C. P. E. & Heckman, N. E. (2014). Switching nonparametric regression models. *Journal of Nonparametric Statistics*, 26, 617–637.

de Souza, C. P. E., Heckman, N. E., & Xu, F. (2017). Switching nonparametric regression models for multi-curve data. *Canadian Journal of Statistics*, 45, 442–460.

DeRuiter, S. L., Langrock, R., Skirbutas, T., Goldbogen, J. A., Calambokidis, J., Friedlaender, A. S., & Southall, B. L. (2017). A multivariate mixed hidden Markov model for blue whale behaviour and responses to sound exposure. *The Annals of Applied Statistics*, 11, 362–392.

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., & Punjabi, N. M. (2009). Multilevel functional principal component analysis. *The Annals of Applied Statistics*, 3, 458–488.

Douc, R., Garivier, A., Moulines, E., & Olsson, J. (2011). Sequential Monte Carlo smoothing for general state space hidden Markov models. *The Annals of Applied Probability*, 21, 2109–2145.

Fehlmann, G., O'Riain, J., Hopkins, P. W., O'Sullivan, J., Holton, M. D., Shepard, E. L. C., & King, A. J. (2017). Identification of behaviours from accelerometer data in a wild social primate. *Animal Biotelemetry*, 5, 6.

Fischer, W. & Meier-Hellstern, K. S. (1993). The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation*, 18, 149–171.

Ford, J. K. B. & Ellis, G. M. (2006). Selective foraging by fish-eating killer whales *Orcinus orca* in British Columbia. *Marine Ecology Progress Series*, 316, 185–199.

Ford, J. K. B., Ellis, G. M., Olesiuk, P. F., & Balcomb, K. C. (2009). Linking killer whale survival and prey abundance: Food limitation in the oceans' apex predator?. *Biology Letters*, 6, 139–142.

Foti, N., Xu, J., Laird, D., & Fox, E. (2014). Stochastic variational inference for hidden Markov models. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., & Weinberger, K. Q. (Eds.) *Advances in Neural Information Processing Systems*, Vol. 27, Curran Associates Inc., Red Hook, NY.

Fu, E. & Heckman, N. (2019). Model-based curve registration via stochastic approximation EM algorithm. *Computational Statistics and Data Analysis*, 131, 159–175.

Gao, Y., Shang, H. L., & Yang, Y. (2019). High-dimensional functional time series forecasting: An application to age-specific mortality rates. *Journal of Multivariate Analysis*, 170, 232–243. (Special issue on Functional Data Analysis and Related Topics)

Getman, A., Cooper, C. D., Key, G., Zhou, H., & Frankle, N. (2009). Detection of mobile machine damage using accelerometer data and prognostic health monitoring techniques. *2009 IEEE Workshop on Computational Intelligence in Vehicles and Vehicular Systems*, IEEE, New York, 101–104.

Gleiss, A. C., Wilson, R. P., & Shepard, E. L. C. (2011). Making overall dynamic body acceleration work: On the theory of acceleration as a proxy for energy expenditure. *Methods in Ecology and Evolution*, 2, 23–33.

Green, J. A., Halsey, L. G., Wilson, R. P., & Frappell, P. B. (2009). Estimating energy expenditure of animals using the accelerometry technique: Activity, inactivity and comparison with the heart-rate technique. *Journal of Experimental Biology*, 212, 745–746.

Hastie, G. D., Rosen, D. A. S., & Trites, A. W. (2006). The influence of depth on a breath-hold diver: Predicting the diving metabolism of Steller sea lions (*Eumetopias jubatus*). *Journal of Experimental Marine Biology and Ecology*, 3, 163–170.

Heerah, K., Woillez, M., Fablet, R., Garren, F., Martin, S., & De Pontual, H. (2017). Coupling spectral analysis and hidden Markov models for the segmentation of behavioural patterns. *Movement Ecology*, 5, 20.

Hooten, M. B., King, R., & Langrock, R. (2017). Guest editor's introduction to the special issue on "Animal movement modeling". *Journal of Agricultural, Biological and Environmental Statistics*, 22, 224–231.

Isojunno, S., Sadykova, D., DeRuiter, S., Curé, C., Visser, F., Thomas, L., Miller, P. J. O., & Harris, C. M. (2017). Individual, ecological, and anthropogenic influences on activity budgets of long-finned pilot whales. *Ecosphere*, 8, e02044.

Jeanniard-du Dot, T., Guinet, C., Arnould, J. P., Speakman, J. R., & Trites, A. W. (2016). Accelerometers can measure total and activity-specific energy expenditures in free-ranging marine mammals only if linked to time-activity budgets. *Functional Ecology*, 31, 377–386.

Jeanniard du Dot, T., Trites, A. W., Arnould, J. P. Y., Speakman, J. R., & Guinet, C. (2016). Activity-specific metabolic rates for diving, transiting, and resting at sea can be estimated from time-activity budgets in free-ranging marine mammals. *Ecology and Evolution*, 7, 2969–2976.

Juang, B. H. & Rabiner, L. R. (1991). Hidden Markov models for speech recognition. *Technometrics*, 33, 251–272.

Kokoszka, P. & Reimherr, M. (2018). *Chapter 8: Introduction to Functional Data Analysis*. Chapman and Hall/CRC Press, Boca Raton, FL.

Langrock, R., Adam, T., Leos-Barajas, V., Mews, S., Miller, D. L., & Papastamatiou, Y. P. (2018). Spline-based nonparametric inference in general state-switching models. *Statistica Neerlandica*, 72, 179–200.

Langrock, R., Swihart, B. J., Caffo, B. S., Punjabi, N. M., & Crainiceanu, C. M. (2013). Combining hidden Markov models for comparing the dynamics of multiple sleep electroencephalograms. *Statistics in Medicine*, 32, 3342–3356.

Lawler, E., Whoriskey, K., Aeberhard, W. H., Field, C., & Mills Flemming, J. (2019). The conditionally autoregressive hidden Markov model (CarHMM): Inferring behavioural states from animal tracking data exhibiting conditional autocorrelation. *Journal of Agricultural, Biological and Environmental Statistics*, 24, 651–668.

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, 232–248.

Liu, Y.-Y., Li, S., Li, F., Song, L., & Rehg, J. M. (2015). Efficient learning of continuous-time hidden Markov models for disease progression. *Advances in Neural Information Processing Systems*, 28, 3599–3607.

Lucero, P., Sánchez, R., Macancela, J., Cabrera, D., Cerrada, M., Li, C., & Alonso, H. R. (2019). Accelerometer placement comparison for crack detection in railway axles using vibration signals and machine learning. *2019 Prognostics and System Health Management Conference (PHM-Paris)*, IEEE, New York, 291–296. https://ieeexplore.ieee.org/document/8756473.

McClintock, B. T., Langrock, R., Gimenez, O., Cam, E., Borchers, D. L., Glennie, R., & Patterson, T. A. (2020). Uncovering ecological state dynamics with hidden Markov models. *Ecology Letters*, 23, 1878–1903.

Michelot, T. & Blackwell, P. G. (2019). State-switching continuous-time correlated random walks. *Methods in Ecology and Evolution*, 10, 637–649.

Morris, J. S., Arroyo, C., Coull, B. A., Ryan, L. M., Herrick, R., & Gortmaker, S. L. (2006). Using wavelet-based functional mixed models to characterize population heterogeneity in accelerometer profiles. *Journal of the American Statistical Association*, 101, 1352–1364.

Noren, D. P. (2011). Estimated field metabolic rates and prey requirements of resident killer whales. *Marine Mammal Science*, 27, 60–77.

Nunes, A. (2019). *Divebomb version 1.0.7*. Ocean Tracking Network. gitlab.oceantrack.org/anunes/divebomb.

Patterson, T. A., Parton, A., Langrock, R., Blackwell, P. G., Thomas, L., & King, R. (2017). Statistical modelling of individual animal movement: An overview of key methods and a discussion of practical challenges. *Advances in Statistical Analysis*, 101, 399–438.

Pohle, J., Langrock, R., van Beest, M., & Schmidt, N. M. (2017). Selecting the number of states in hidden Markov models: Pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics*, 22, 1–24.

Qasem, L., Cardew, A., Wilson, A., Griffiths, I., Halsey, L. G., Shepard, E. L. C., Gleiss, A. C., & Wilson, R. (2012). Tri-axial dynamic acceleration as a proxy for animal energy expenditure; should we be summing values or calculating the vector? *PLoS One*, 7, e31187.

Ramsay, J. O. & Silverman, B. W. (2005). *Functional Data Analysis*, 2nd ed., Springer, New York City, NY.

Rice, J. A. & Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, 57, 253–259.

Scott, S. L. (2002). Bayesian methods for hidden Markov models. *Journal of the American Statistical Association*, 97, 337–351.

Shorter, K. A., Shao, Y., Ojeda, L., Barton, K., Rocho-Levine, J., van der Hoop, J., & Moore, M. (2017). A day in the life of a dolphin: Using bio-logging tags for improved animal health and well-being. *Marine Mammal Science*, 33, 785–802.

Simon, M., Johnson, M., & Madsen, P. T. (2012). Keeping momentum with a mouthful of water: Behavior and kinematics of humpback whale lunge feeding. *Journal of Experimental Biology*, 215, 3786–3798.

Sutherland, W. J. (1998). The importance of behavioural studies in conservation biology. *Animal Behaviour*, 56, 801–809.

Tennessen, J., Holt, M. M., Ward, E. J., Hanson, M. B., Emmons, C. K., Giles, D. A., & Hogan, J. T. (2019a). Hidden Markov models reveal temporal patterns and sex differences in killer whale behavior. *Scientific Reports*, 9, 14951.

Tennessen, J. B., Holt, M. M., Hanson, M. B., Emmons, C. K., Giles, D. A., & Hogan, J. T. (2019b). Kinematic signatures of prey capture from archival tags reveal sex differences in killer whale foraging activity. *Journal of Experimental Biology*, 222.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., et al. (2019). *SciPy 1.0—Fundamental Algorithms for Scientific Computing in Python*. arXiv e-prints arXiv:1907.10121.

Whoriskey, K., Auger-Méthé, M., Albertsen, C. M., Whoriskey, F. G., Binder, T. R., Krueger, C. C., & Mills Flemming, J. (2016). A hidden Markov movement model for rapidly identifying behavioral states from animal tracks. *Ecology and Evolution*, 7, 2112–2121.

Williams, R. & Noren, D. P. (2009). Swimming speed, respiration rate, and estimated cost of transport in adult killer whales. *Marine Mammal Science*, 25, 327–350.

Williams, T. M., Haun, J. E., & Friedl, W. A. (1999). The diving physiology of bottlenose dolphins (*Tursiops truncatus*): I. Balancing the demands of exercise for energy conservation at depth. *Journal of Experimental Biology*, 202, 2739–2748.

Wilson, R. P., Börger, L., Holton, M. D., Scantlebury, D. M., Gómez-Laich, A., Quintana, F., Rosell, F., et al. (2019). Estimates for energy expenditure in free-living animals using acceleration proxies: A reappraisal. *Journal of Animal Ecology*, 89, 161–172.

Wright, B. M., Ford, J. K. B., Ellis, G. M., Deecke, V. B., Shapiro, A. D., Battaile, B. C., & Trites, A. W. (2017). Fine-scale foraging movements by fish-eating killer whales (*Orcinus orca*) relate to the vertical distributions and escape responses of salmonid prey (*Oncorhynchus* spp.). *Movement Ecology*, 5.

Xin, G., Hamzaoui, N., & Antoni, J. (2018). Semi-automated diagnosis of bearing faults based on a hidden Markov model of the vibration signals. *Measurement*, 127, 141–166.

Xu, Z., Laber, E. B., & Staicu, A.-M. (2020). *Hierarchical Continuous Time Hidden Markov Model, with Application in Zero-Inflated Accelerometer Data*. Springer International Publishing, Cham.

Yao, F., Müller, H.-G., & Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100, 577–590.

Zucchini, W., Macdonald, I. L., & Langrock, R. (2016). *Hidden Markov Models for Time Series: An Introduction Using R*, 2nd ed., CRC Press, Boca Raton, FL.

APPENDIX

Detailed Description of the Data Simulation in Section 4.1

We can easily simulate realizations of the coarse-scale HMM (X and Y) given the parameters Γ and θ. For each dive t, we can also easily generate the fine-scale hidden Markov chain $\tilde{X}^*_t = \{\tilde{X}^*_{t,1}, \ldots, \tilde{X}^*_{t,\tilde{T}^*_t}\}$ according to one of the transition probability matrices $\Gamma^{*(1)}$ or $\Gamma^{*(2)}$ depending upon the value of X_t. This determines the sequence of fine-scale hidden states corresponding to each window. Recall that the fine-scale model is based on a sequence of $\tilde{T}^*_t$ 2-s windows, each containing 100 observations, and that our model is formulated in terms of quantities derived from the raw data within each window (namely the average acceleration, $\tilde{A}^*_{t,\tilde{t}^*}$, and wiggliness, $\tilde{W}^*_{t,\tilde{t}^*}$). Generating the raw acceleration data from $\tilde{A}^*_{t,\tilde{t}^*}$ and $\tilde{W}^*_{t,\tilde{t}^*}$ is not straightforward. In our simulation study, we generate raw acceleration data so that we can visualize our results in terms of the underlying data curves for each dive. Here, we explain how

we generate the acceleration curves so that $\tilde{A}^*_{t,\tilde{t}^*}$ and $\tilde{W}^*_{t,\tilde{t}^*}$ both follow the specified model. A key component is the DFT of the 100 raw acceleration values in window $\tilde{t}^*$ of dive t,

$$\hat{Y}^{*(k)}_{t,\tilde{t}^*} = \text{DFT}\left\{ Y^*_{t,100(\tilde{t}^*-1)+1}, \cdots, Y^*_{t,100\tilde{t}^*} \right\}(k)$$

for $k \geq 0$, as defined in Equation (6).

We simulate the raw acceleration data for dive t in three steps: (1) simulate the average acceleration within each window $\left(\hat{Y}^{*(0)}_{t,\tilde{t}^*}\right)$; (2) simulate all other Fourier coefficients within each window $\left(\hat{Y}^{*(k)}_{t,\tilde{t}^*}, k = 1, \ldots, 99\right)$; and (3) take the inverse DFT of $\hat{Y}^*_{t,\tilde{t}^*}$,

$$\left\{ Y^*_{t,100(\tilde{t}^*-1)+1}, \ldots, Y^*_{t,100\tilde{t}^*} \right\} = \text{IDFT}\left\{ \hat{Y}^{*(0)}_{t,\tilde{t}^*}, \ldots, \hat{Y}^{*(99)}_{t,\tilde{t}^*} \right\}$$

for $\tilde{t}^* = 1, \ldots, \tilde{T}^*_t$. The details of steps (1) and (2) are given below.

For step (1), we generate $\hat{Y}^{*(0)}_{t,1}, \ldots, \hat{Y}^{*(0)}_{t,\tilde{T}^*_t}$ as a CarHMM with the underlying Markov state sequence $\tilde{X}^*_{t,1}, \ldots, \tilde{X}^*_{t,\tilde{T}^*_t}$ and a random first emission. Specifically, we let

$$\hat{Y}^{*(0)}_{t,1} \mid \tilde{X}^*_{t,1} = i^* \;\sim\; \mathcal{N}\left(100\mu_A^{*(\cdot,i^*)}, \left(100\sigma_A^{*(\cdot,i^*)}\right)^2\right)$$

and

$$\hat{Y}^{*(0)}_{t,\tilde{t}^*} \mid \tilde{X}^*_{t,\tilde{t}^*} = i^*, \hat{Y}^{*(0)}_{t,\tilde{t}^*-1} \;\sim\; \mathcal{N}\left(\phi_A^{*(\cdot,i^*)}\hat{Y}^{*(0)}_{t,\tilde{t}^*-1} + 100\left(1-\phi_A^{*(\cdot,i^*)}\right)\mu_A^{*(\cdot,i^*)}, \left(100\sigma_A^{*(\cdot,i^*)}\right)^2\right), \tag{A1}$$

for $\tilde{t}^* = 2, \ldots, \tilde{T}^*_t$, where $\mu_A^{*(\cdot,i^*)}$, $\sigma_A^{*(\cdot,i^*)}$, and $\phi_A^{*(\cdot,i^*)}$ are as defined in the simulation study in Section 4.

For step (2), we first construct $\hat{Y}^{*(k)}_{t,\tilde{t}^*}$ for $k = 1, \ldots, 49$, as

$$\hat{Y}^{*(k)}_{t,\tilde{t}^*} = a^{(k)}_{t,\tilde{t}^*} i \sqrt{b^{(k)}_{t,\tilde{t}^*}}, \tag{A2}$$

where the $a^{(k)}_{t,\tilde{t}^*}$s are independent and equal to either 1 or -1 each with probability 1/2 and i is the imaginary unit. We include i in Equation (A2) to force all variation within a window to take the form of a sine wave, which reduces the variation between the end points of windows relative to a cosine wave. Given the fine scale states, the $b^{(k)}_{t,\tilde{t}^*}$s are independent of one another and independent of $a^{(k)}_{t,\tilde{t}^*}$s. The distribution of $b^{(k)}_{t,\tilde{t}^*}$ is

$$\begin{aligned}
b^{(k)}_{t,\tilde{t}^*} \mid \tilde{X}^*_{t,\tilde{t}^*} = 1 \;&\sim\; \text{Gamma}(11.03/(k+1)^3, 15.54) \\
b^{(k)}_{t,\tilde{t}^*} \mid \tilde{X}^*_{t,\tilde{t}^*} = 2 \;&\sim\; \text{Gamma}(4.80/(k+1)^3, 515.38) \\
b^{(k)}_{t,\tilde{t}^*} \mid \tilde{X}^*_{t,\tilde{t}^*} = 3 \;&\sim\; \text{Gamma}(2.31/(k+1)^3, 20{,}023.44).
\end{aligned} \tag{A3}$$

The first argument of Gamma$(\cdot,\cdot)$ is the shape parameter and the second is the scale parameter. The squared magnitude of the kth Fourier coefficient is equal to $b^{(k)}_{t,\tilde{t}^*}$, which decays on the order of $1/k^3$ to "smooth out" the raw acceleration data.

We then define the remaining 50 Fourier coefficients as $\hat{Y}^{*(50)}_{t,\tilde{t}^*} = 0$ and $\hat{Y}^{*(k)}_{t,\tilde{t}^*} = -\hat{Y}^{*(100-k)}_{t,\tilde{t}^*}$ for $k = 51, \ldots, 99$. This guarantees that the inverse DFT is real-valued.

We now show that this construction of the raw acceleration data results in the distributions listed in Section 4.1. It suffices to show that the construction of the DFTs, the $\hat{Y}^{(k)}_{t,\tilde{t}^*}$s, yields the desired distributions.

First, since $\hat{Y}^{(0)}_{t,\tilde{t}^*} = \sum_{n=1}^{100} Y^*_{t,100(\tilde{t}^*-1)+n} = 100\tilde{A}^*_{t,\tilde{t}^*}$, Equation (A1) implies that $\tilde{A}^*_{t,1}, \ldots, \tilde{A}^*_{t,\tilde{T}^*}$ follows a CarHMM with normal emission distributions and parameters as defined in the simulation study in Section 4.

From Equations (7) and (A2), the "wiggliness" within window $\tilde{t}^*$ of dive t is

$$\tilde{W}^*_{t,\tilde{t}^*} = \sum_{k=1}^{\tilde{\omega}} \left|\left|\hat{Y}^{(k)}_{t,\tilde{t}^*}\right|\right|^2 = \sum_{k=1}^{\tilde{\omega}} b^{(k)}_{t,\tilde{t}^*}.$$

If $\tilde{\omega} < 50$, then $\tilde{W}^*_{t,\tilde{t}^*}$ is the sum of independent, gamma-distributed random variables with identical scale parameters, so the distribution of $\tilde{W}^*_{t,\tilde{t}^*}$ is also gamma-distributed. Thus, by Equation (A3)

$$\tilde{W}^*_{t,\tilde{t}^*} \mid \tilde{X}^*_{t,\tilde{t}^*} = 1 \;\sim\; \text{Gamma}\left(\sum_{k=1}^{\tilde{\omega}} 11.03/(k+1)^3, 15.54\right),$$

$$\tilde{W}^*_{t,\tilde{t}^*} \mid \tilde{X}^*_{t,\tilde{t}^*} = 2 \;\sim\; \text{Gamma}\left(\sum_{k=1}^{\tilde{\omega}} 4.80/(k+1)^3, 515.38\right),$$

and

$$\tilde{W}^*_{t,\tilde{t}^*} \mid \tilde{X}^*_{t,\tilde{t}^*} = 3 \;\sim\; \text{Gamma}\left(\sum_{k=1}^{\tilde{\omega}} 2.31/(k+1)^3, 20{,}023.44\right).$$

Setting $\tilde{\omega}$ to 10 and carrying out simple calculation of the mean and variance of a gamma distribution gives the desired values for $\mu_W^{*(\cdot,i^*)}$ and $\sigma_W^{*(\cdot,i^*)}$.

Likelihood of the CarHHMM-DFT

The overall likelihood of the CarHHMM-DFT is

$$\mathcal{L}_{\text{CarHHMM-DFT}}(\theta, \theta^*, \Gamma, \Gamma^*; y, \tilde{y}^*) = \delta P\left(y_1, \tilde{y}^*_1; \theta, \theta^*, \Gamma^*\right) \prod_{t=2}^{T} \Gamma P\left(y_t, \tilde{y}^*_t; \theta, \theta^*, \Gamma^*\right) \mathbf{1}_N,$$

where

$$P\left(y_t, \tilde{y}^*_t; \theta, \theta^*, \Gamma^*\right) = \text{diag}\Big[f^{(1)}\left(y_t; \theta^{(1)}\right) \mathcal{L}_{\text{fine}}\left(\theta^*, \Gamma^{*(1)}; \tilde{y}^*_t\right), \ldots, f^{(N)}\left(y_t; \theta^{(N)}\right) \mathcal{L}_{\text{fine}}\left(\theta^*, \Gamma^{*(N)}; \tilde{y}^*_t\right)\Big]$$

and $f^{(i)}\left(y_t; \theta^{(i)}\right)$ is the emission distribution of dive duration given that $X_t = i$. The likelihood corresponding to the fine-scale model is

$$\mathcal{L}_{\text{fine}}\left(\theta^*, \Gamma^{*(i)}; \tilde{y}^*_t\right) = \delta^{*(i)} \prod_{\tilde{t}^*=2}^{\tilde{T}^*_t} \Gamma^{*(i)} P\left(\tilde{y}^*_{t,\tilde{t}^*} \mid \tilde{y}^*_{t,\tilde{t}^*-1}; \theta^*\right) \mathbf{1}_{N^*},$$

where $P\left(\tilde{y}^*_{t,\tilde{t}^*} \mid \tilde{y}^*_{t,\tilde{t}^*-1}; \theta^*\right)$ is an $N^* \times N^*$ diagonal matrix with the (i^*, i^*)th entry $f^{*(\cdot,i^*)}\left(\tilde{y}^*_{t,\tilde{t}^*} \mid \tilde{y}^*_{t,\tilde{t}^*-1}; \theta^{*(\cdot,i^*)}\right)$. Recall that $f^{*(\cdot,i^*)}\left(\cdot \mid \tilde{y}^*_{t,t^*-1}; \theta^{*(\cdot,i^*)}\right)$ is the probability density function of $\tilde{Y}^*_{t,\tilde{t}^*}$ when $\tilde{X}^*_{t,\tilde{t}^*} = i^*$ and $\tilde{Y}^*_{t,\tilde{t}^*-1} = \tilde{y}^*_{t,\tilde{t}^*-1}$.

Received 14 January 2021
Accepted 14 September 2021