**Appendix S5.** Posterior predictive checks.

Assessing the goodness-of-fit of hidden state models fitted in a Bayesian framework is less straightforward than in comparable frequentist models (Jonsen et al., 2013). Alternatively, posterior predictive checks can be used to assess the ability of the model to replicate various features of the observed data. De Haan-Rietdijk et al. (2017), Morales, Haydon, Frair, Holsinger, & Fryxell (2004), and Shirley, Small, Lynch, Maisto, & Oslin (2010) proposed several procedures to this purpose, which all aim to evaluate whether the simulation of new data using the posterior estimates of model parameters can generate summary statistics that are comparable with the same summary statistics in the empirical dataset.

We followed the procedure described in Appendix S4 to simulate 1,000 new datasets of comparable size based on random draws from the posterior distribution of the state-dependent parameters of the emission distributions and transition probabilities. By doing so, we implicitly accounted for uncertainty in these parameters. Results of the simulations were then compared to the dataset of filtered and regularised locations that was used to fit the model.

First, we compared the activity budget, i.e., the proportion of time spent in each latent state across all segments and tracks. The proportion of time spent in each state in the original data was plotted over the distribution of proportions obtained in the simulated datasets. Similarly, we calculated the mean and standard deviation of the duration of stays (i.e., the number of consecutive one-minute steps) in each state for each simulated dataset, and compared their distribution with the same statistics in the original data. We then considered the number of behavioural transitions per segment, corrected by the length of each segment. As for the previous metric, we plotted the mean and standard deviation in the original dataset over the distribution of the same statistics in the simulated dataset. Because the distributions of relative number of behavioural transitions were not Gaussian and, therefore, the mean and standard deviation did not necessarily provide appropriate summaries as for the other metrics, we also plotted them together to verify their degree of overlap. Finally, we plotted the autocorrelation function plot (ACF) for all response variables in the original and simulated data.

The activity budget estimated from the simulated data was, overall, satisfactorily comparable with the activity budget in the original data (Fig. S4a). In the simulated tracks, eagles allocated more time than expected to state 4 (on the ground or perching), which also partially affected the proportional occurrence of other states. However, this could result from the fact that, at night or when not flying (that is, when an eagle is most likely to spend time in this state), the GPS device was programmed to record location data less frequently, resulting in a higher chance of gap occurrence; on the other hand, gaps were distributed randomly in the simulated data. This is also reflected in the higher mean duration of stays in this state in the simulated data (Fig. S4b). For all other states, the mean duration of stays in the original data was higher than in the simulated data, highlighting the potential problems with the validity of the Markov property for the short time interval that are discussed in the main text (Fig. S4b and c). Similarly, the relative number of behavioural transitions per segment was slightly higher in the simulated data (Fig. S4d). As demonstrated in the overlap between the two distributions, this reflects a lower occurrence of intermediate numbers of transitions, which could characterise longer stretches of tracks spent in given states. Finally, the ACF plots confirmed these patterns and highlighted the occurrence of residual autocorrelation in some response variables (particularly step length and hierarchical slope position) under specific states (Fig. S5).

**References**

de Haan-Rietdijk, S., Kuppens, P., Bergeman, C. S., Sheeber, L. B., Allen, N. B., & Hamaker, E. L. (2017). On the use of mixed Markov models for intensive longitudinal data. *Multivariate Behavioral Research*, *52*(6), 747–767. doi:10.1080/00273171.2017.1370364

Jonsen, I., Basson, M., Bestley, S., Bravington, M. V., Patterson, T. A., Pedersen, M. W., … Wotherspoon, S. J. (2013). State-space models for bio-loggers: A methodological road map. *Deep Sea Research Part II: Topical Studies in Oceanography*, *88*–*89*, 34–46. doi:10.1016/j.dsr2.2012.07.008

Morales, J. M., Haydon, D. T., Frair, J., Holsinger, K. E., & Fryxell, J. M. (2004). Extracting more out of relocation data: building movement models as mixtures of random walks. *Ecology*, *85*(9), 2436–2445. doi:10.1890/03-0269

Shirley, K. E., Small, D. S., Lynch, K. G., Maisto, S. A., & Oslin, D. W. (2010). Hidden Markov models for alcoholism treatment trial data. *The Annals of Applied Statistics*, *4*(1), 366–395. doi:10.1214/09-AOAS282