Received: 20 October 2023 | Accepted: 17 April 2024

DOI: 10.1111/2041-210X.14336

Methods in Ecology and Evolution

RESEARCH ARTICLE

Using lineups to evaluate goodness of fit of animal movement models

John Fieberg[1] | Smith Freeman[1] | Johannes Signer[2]

[1]Department of Fisheries, Wildlife, and Conservation Biology, University of Minnesota, St. Paul, Minnesota, USA

[2]Wildlife Sciences, Faculty of Forestry and Forest Ecology, University of Göttingen, Göttingen, Germany

Correspondence
John Fieberg
Email: jfieberg@umn.edu

Funding information
National Aeronautics and Space Administration, Grant/Award Number: 80NSSC21K1182; Minnesota Agricultural Experimental Station

Handling Editor: Theoni Photopoulou

Abstract

1. Movement models are frequently fit to animal location data to understand how individuals respond to and interact with local environmental features. Several open-source software packages are available for analysing animal movements and can facilitate parameter estimation, yet there are relatively few methods available for evaluating model goodness of fit.
2. We describe how a simple graphical technique, the *lineup protocol*, can be used to evaluate goodness of fit of integrated step-selection analyses and hidden Markov models, but the method can be applied much more broadly. We leverage the ability to simulate data from fitted models and demonstrate the approach using both an integrated step-selection analysis and a hidden Markov model applied to fisher (*Pekania pennanti*) data.
3. A variety of responses and movement metrics can be used to evaluate models, and the lineup protocol can be tailored to focus on specific model assumptions or movement features that are of primary interest. Although it is possible to evaluate statistical significance using a formal hypothesis test, the method can also be used in a more exploratory fashion (e.g. to explore variability in model behaviour across stochastic simulations or to identify areas where the model could be improved).
4. We provide coded examples and vignettes to demonstrate the flexibility of the approach. We encourage movement ecologists to consider how their models will be applied when choosing appropriate graphical responses for evaluating goodness of fit.

KEYWORDS
animal movement, assumptions, goodness of fit, hidden Markov model, integrated step-selection analysis, lineup, simulation, telemetry

1 | INTRODUCTION

Technological advances, including smaller and better tracking devices (Kays et al., 2015), have led to an exponential increase in animal location data, and have fueled the development of new statistical methods and software for modelling animal movement (Hooten et al., 2017; Joo et al., 2020). Integrated step-selection analyses (ISSAs) (Avgar et al., 2016; Fieberg et al., 2021) and hidden Markov

models (HMMs) (Langrock et al., 2012; McClintock et al., 2012), in particular, are extremely popular for analysing wildlife telemetry data due to the availability of open-source software (`amt` and `momentuHMM`) for implementing them (McClintock & Michelot, 2018; Signer et al., 2019). These approaches, as well as recently developed methods that make it possible to combine the two frameworks (Klappstein et al., 2023; Pohle et al., 2024), allow researchers to fit rich models to telemetry data in which movements may vary spatially and temporally as a function of environmental features (e.g. land-cover types, distance to roads) and latent (unobserved) behavioural states (e.g. representing whether an animal is foraging or travelling).

Despite the popularity of these analytical frameworks, there are relatively few methods available for evaluating their goodness of fit. The `momentuHMM` package returns pseudo-residuals from fitted HMMs, which can be used to evaluate fit (e.g. using a quantile-quantile plot) or to explore potential violations of important assumptions (e.g. lack of autocorrelation in the residuals). Methods for validating ISSAs are almost non-existent, likely due to the 'tricks' commonly used for inference; for example, a two-step procedure is typically used to estimate movement parameters via importance sampling (Avgar et al., 2016; Fieberg et al., 2021; Michelot et al., 2024), and random-effect specifications typically rely on a likelihood equivalence between Poisson regression with stratum-specific intercepts and conditional logistic regression models (Chatterjee et al., 2024; Muff et al., 2020). One approach to model evaluation for ISSAs, recently implemented in the `amt` package and referred to as *used-habitat-calibration plots* (Fieberg et al., 2018), focuses on whether the distribution of environmental characteristics (e.g. elevation, land-cover class) at model-predicted locations matches the distribution of those same characteristics measured at locations visited by the individual(s). This method relies on predictive simulation techniques to derive a simulation envelope for the distribution of environmental conditions at used locations, and shares similarities with Bayesian approaches that evaluate goodness of fit using posterior-predictive distributions (Conn et al., 2018).

Here, we develop a general and flexible approach for evaluating goodness of fit of ISSAs and HMMs, and animal movement models in general, by leveraging our ability to simulate trajectories from fitted models. Whereas Fieberg et al. (2018) derived predictive distributions using cross-validation at the step level (i.e. predicted locations were conditioned on the previously observed location), we will consider simulations of multi-step trajectories. We can then consider how these trajectories compare to observed movements using a wide range of statistical and graphical summaries. A rigorous goodness-of-fit test can be implemented using the lineup protocol, a formal graphical null hypothesis testing framework (Buja et al., 2009; Wickham et al., 2010). Alternatively, we can use the visualizations in an exploratory way to better understand the diversity of movements that may be generated from fitted models or to determine if our models are capable of generating important behaviours exhibited by our study animals. We discuss how these approaches can compliment the suite of tools currently available for evaluating goodness of fit.

2 | THE LINEUP PROTOCOL

As with any goodness-of-fit test, we will assume the following null and alternative hypotheses:

H_0: The data are consistent with the assumed model.
H_A: The data are not consistent with the assumed model.

Although the lineup protocol takes its name from police lineups, the method actually depends on the power of the human eye to decipher patterns rather than to identify a known entity among unknown entities (Wickham et al., 2010). In short, we will simulate several data sets from our fitted model, summarize the simulated and observed (i.e. real) data sets graphically, and see if the real data are easily decipherable from the simulated data sets. If they are, then we will reject the null hypothesis and conclude that our model is missing one or more important features (i.e. the assumptions of the model are not met). Otherwise, we will fail to reject the null hypothesis and conclude that the model is capable of capturing the characteristics of the observed data (i.e. the model assumptions are plausible).

When using a lineup, a plot serves as our test statistic, and plots of the simulated data collectively serve as our sampling distribution under the null hypothesis. In our examples, we will generate plots with 20 panels visualizing the observed data and 19 simulated data sets, with the observed-data panel randomly located (e.g. Figure 1). If the real data are consistent with the assumed model (i.e. if the null hypothesis is true), then we would have a 1 in 20 (or 0.05) chance of selecting any one of the panels. Thus, selecting the panel with the observed data implies that the p-value for the hypothesis test is ≤ 0.05; the $\leq$ reflects the fact that the precision with which we are able to calculate the *p*-value is limited by the number of panels included in the plot (we can increase precision by including more panels). To increase the statistical power of the test, one can offer lineups to multiple observers. If each observer is given a unique lineup (generated using a different subset of the real data and new simulated data sets), then the observers' choices can be treated as independent. In this case, the p-value for the goodness-of-fit test can be determined using the cumulative distribution function of the binomial distribution with $N =$ the number of lineups/observers and $p = 1/n$ (where n is the number of panels in each lineup). For example, if a unique lineup with 20 plots were shown to $N = 10$ independent observers, and 3 identified the panel containing the real data, the *p*-value would be given by $\sum_{i=3}^{10} \binom{10}{i} 0.05^i (1-0.05)^i$ and could be calculated in R using `sum(dbinom(3:10, size=10, p=0.05))` $=0.012$.

3 | INTEGRATED STEP-SELECTION ANALYSIS (ISSA) OF FISHER DATA

3.1 | Methods

We begin by considering an ISSA using tracking data from three fishers (*Pekania pennanti*) available through Movebank (Brown et al., 2012; LaPoint et al., 2013a, 2013b). We resampled the data

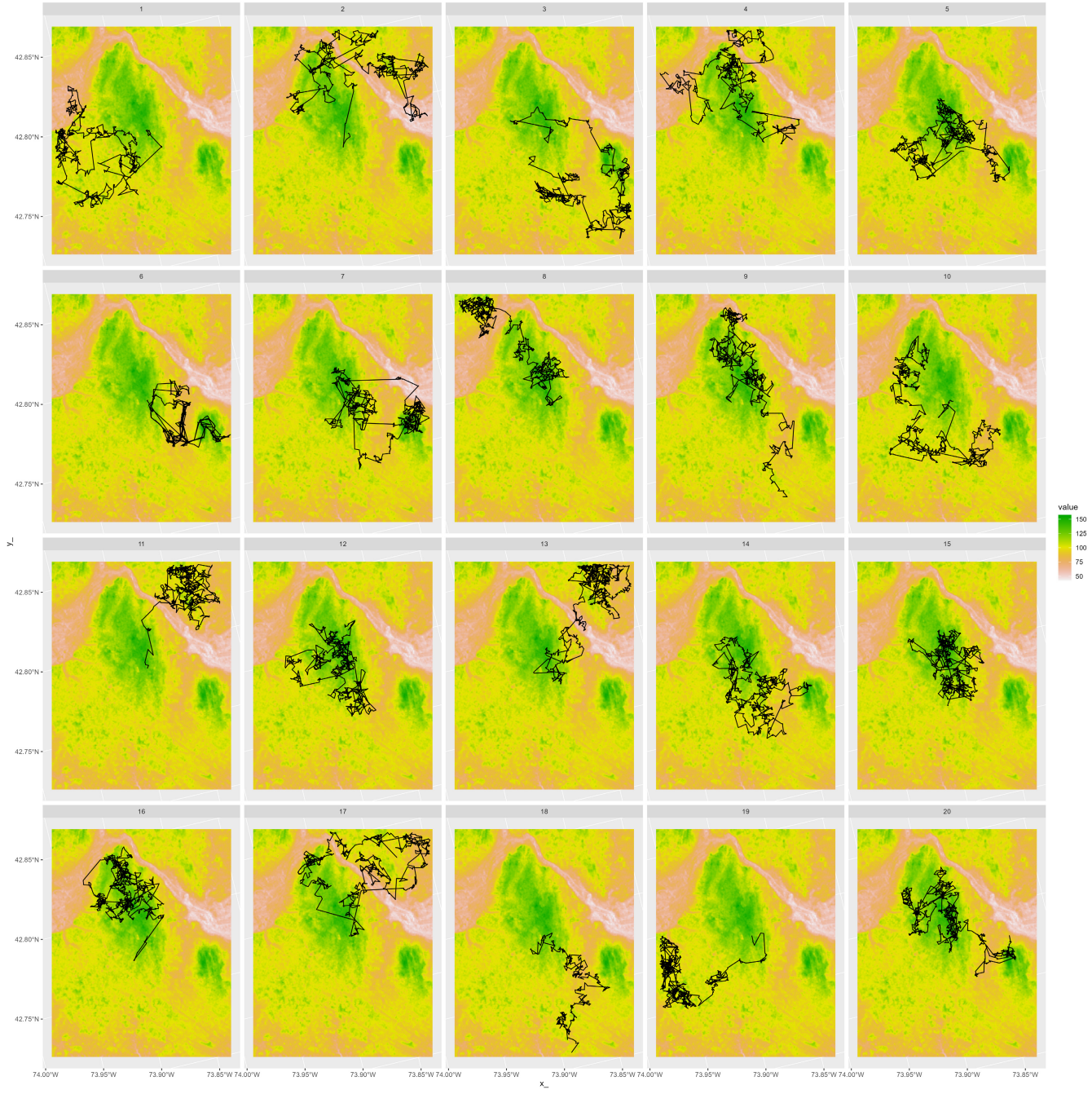

FIGURE 1 Simulated trajectories overlayed on a map of elevation. Nineteen of the panels were generated by simulating observations from a model of animal movement. The model was parameterized using an integrated step-selection analysis with fisher data from LaPoint et al. (2013a, 2013b). The other panel contains an observed trajectory for an individual from the same study but whose data were not used to parameterize the model.

to a fix interval of 10 min with a tolerance of 1 min using the `track_resample()` function in the `R`-package `amt` (Signer et al., 2019). We pooled data from two of the individuals and set aside data from the third individual for model evaluation.

We assumed that the distributions of step lengths and turn angles followed gamma and von Mises distributions, respectively. We estimated tentative parameters of these distributions using the empirical steps connecting consecutive locations, from which we then generated 50 random steps for each observed step using the `random_steps()` function from the `amt` package. We annotated the tracks with environmental variables measuring population density (Center for International Earth Science Information Network (CIESIN) Columbia University, & CIAT, Centro Internacional de Agricultura Tropical, 2005), elevation (U.S./Japan ASTER Science Team, 2009), and landcover class (Defourny et al., 2009) at the end of each simulated and observed step. The original landcover data associated with deciduous, coniferous, or mixed forest were grouped to form a single *forest* indicator variable. We formed strata

by matching each observed step to its 50 associated random steps and fit a conditional logistic regression model with forest, population density and elevation included as covariates. We also included step length, log(step length) and cos(turn angle) in the model to update the movement parameters after accounting for habitat selection (Avgar et al., 2016; Fieberg et al., 2021).

To evaluate the model, we first constructed used-habitat-calibration plots (Fieberg et al., 2018) using the `prep_uhc()` function in the `amt` package (Signer et al., 2019). This approach evaluates the model's ability to predict the environmental characteristics associated with one-step-ahead predictions. For each step in the withheld data set, we used the `prep_uhc()` function to randomly choose between the observed step and the 50 random steps generated from the tentative movement kernel estimated when fitting the ISSA (note, this approach is consistent with Fieberg et al., 2018, but we could instead choose to use the updated movement kernel to generate the random steps); probability weights were assigned to each location (observed and random) using the habitat-selection parameters estimated when fitting the ISSA along with the environmental covariates (population density, elevation, and forest indicator variable) measured at the end of each step. This process was then repeated multiple times to create 95% simulation envelopes for the distribution of the environmental covariates (elevation, population density and forest), which we then compared to the observed distribution of these covariates in the withheld data set (Figure 2a–c).

We also simulated multi-step trajectories starting at the same initial location as the third fisher (i.e. the fisher that was not used to parameterize the model) using the `redistribution_kernel()` and `simulate_path()` functions in the `amt` package and the parameters from the fitted model (Signer et al., 2023). We simulated 19 data sets over the same time interval as the observed data. We matched the pattern of missingness in the observed and simulated trajectories by only including steps with matching timestamps in both the observed and simulated data sets. Rather than create simulation envelopes from the trajectories (as we did with the one-step-ahead predictions), we overlayed distributions of the environmental covariates for both the real and 19 simulated trajectories on the same plot (Figure 2d–f).

We then created a series of lineups using the real and simulated trajectories. In the main text, we consider a lineup showing plots of the trajectories overlayed on a map of elevation (Figure 1). In the appendix, we also consider lineups depicting (1) histograms of the elevations at the locations visited by the real and simulated animals (Figure S1), (2) histograms of the distribution of turn angles (Figure S2); (3) scatterplots of step lengths and turn angles (Figure S3) and (4) circular boxplots showing the distribution of turn angles for different quantiles of the step-length distribution (Figure S4). The first three plots were constructed using functions in the `ggplot2` package (Wickham, 2016). The latter two plots were constructed using the `plot_joint_scat()` and `plot_joint_box()` functions in the `cylcop` package (Hodel & Fieberg, 2021, 2022), respectively. In addition, we used the `patchwork` package (Pedersen, 2020) to stitch together the scatterplots into a single multi-panel plot, the `ggtext` package (Wilke & Wiernik, 2022) to annotate the plot, and the `nullabor` package (Buja et al., 2009; Wickham et al., 2010) for helper functions used to create the lineups. A key for identifying the real data in each figure is provided in the appendix.

3.2 | Results

The used-habitat-calibration plots suggest the model does reasonably well at predicting the distribution of environmental covariates for the one-step-ahead predictions (Figure 2a–c). Yet, there is not a lot of variability in the predicted distributions across the different simulations (i.e. the simulation envelopes are fairly narrow). This feature likely reflects the limited variability in the environmental features within the region of space that can be easily accessed between successive locations. Thus, by focusing on one-step-ahead predictions, we may be offering a fairly weak test of the model's predictive capabilities. On the other hand, we see that the model predicts that the third fisher will spend more time in the forest than it actually does (Figure 2c). In addition, the observed distribution of elevations in the test data set falls outside of the simulation envelope for elevation values between roughly 90 and 100 m. Because the simulation envelope is constructed *pointwise* (i.e. using quantiles across a range of elevation values), it would not be valid to formulate and reject a goodness-of-fit test using an $\alpha = 0.05$ significance level based on this visual comparison (Baddeley et al., 2014; Loosmore & Ford, 2006). To avoid multiple testing issues, we would need to pick a particular elevation a priori on which to focus, use methods to create a global rather than pointwise envelope, or develop a test using a single summary statistic from the density curves (e.g. the maximum deviation from the average of all trajectories) (Baddeley et al., 2014).

When we consider the multi-step trajectories (Figure 2a–c), we see that the distribution of environmental covariates is broader and more variable across different replicates. This variability is helpful for understanding the range of possible outcomes that we might expect based on the fitted model. Although we again focused on the environmental features at visited locations, similar plots could be constructed for other movement descriptors (e.g. step lengths and turn angles). We can convey the same information in a lineup where each trajectory is summarized in a separate panel of a multi-panel plot (see e.g. Figure S1 for a lineup showing the distribution of elevations across the real and 19 simulated trajectories). This offers two advantages. First, we can perform a valid goodness-of-fit test based on whether observers are able to identify the real data in the lineup. Second, we can consider much more interesting and flexible visualizations since the information from the different trajectories do not have to be conveyed in a single plot. For example, lineups showing simulated trajectories on spatial maps (Figure 1) make it possible to quickly check whether the model is capable of capturing various features of the animal's movements (e.g. whether they are constrained by linear features). This information would be difficult to visualize in a single plot containing all trajectories.

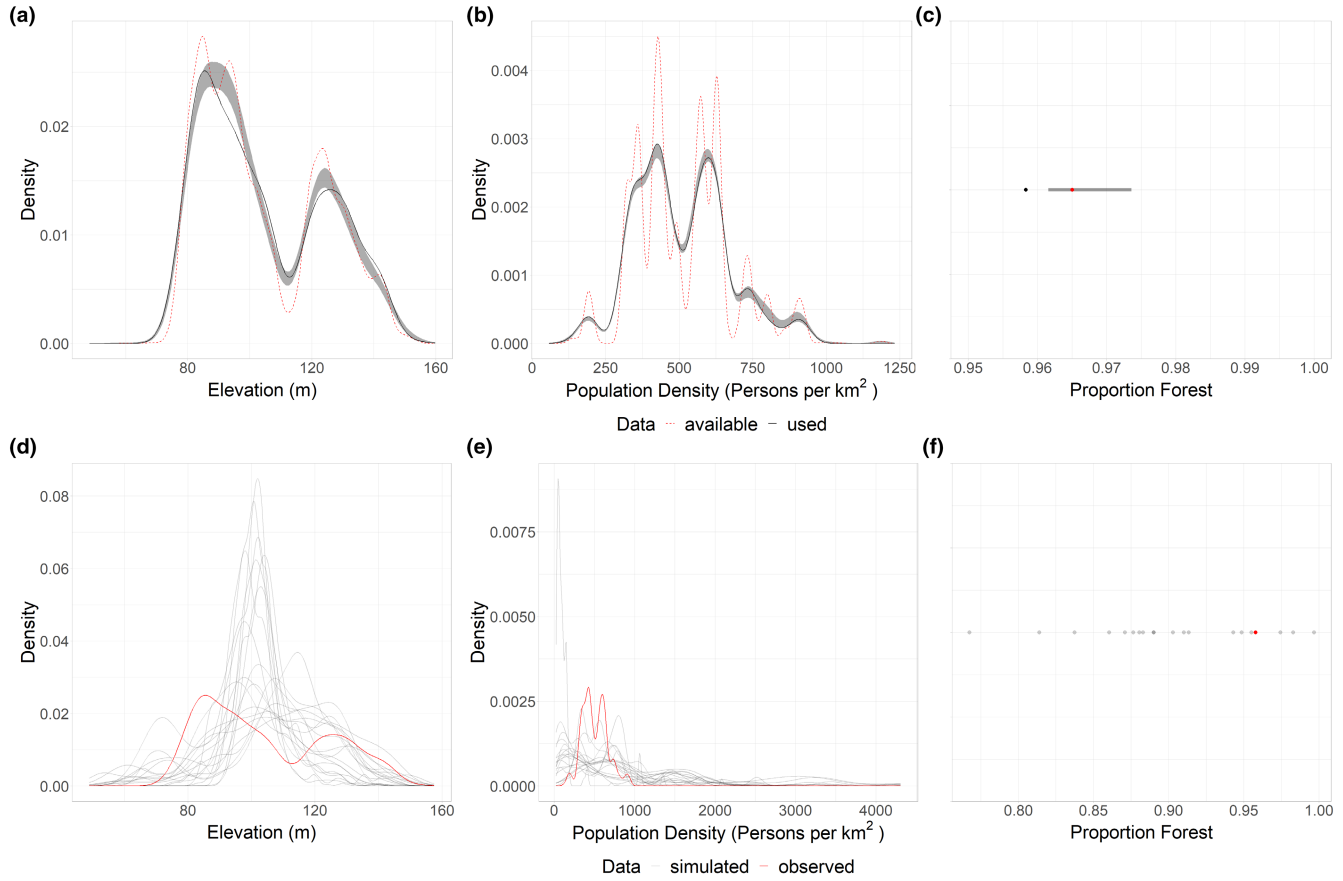

FIGURE 2 Diagnostic plots for the integrated step-selection analysis (ISSA): Used-habitat-calibration plots (a–c) for evaluating one-step ahead predictions and simulations of multi-step trajectories (d–f). Used-habitat-calibration plots show 95% simulation envelopes and were created using the `prep_uhc()` function in the `amt` package (Signer et al., 2019). Multi-step trajectories were created using the `redistribution_kernel()` and `simulate_path()` functions in the `amt` package (Signer et al., 2023).

4 | HIDDEN MARKOV MODEL FIT TO FISHER DATA

4.1 | Methods

We fit a two-state HMM to the same resampled data that we used in the ISSA but ignored the covariates. We represented each 'burst' of steps containing successive locations that were equally spaced in time as separate 'individuals' and removed 'bursts' that contained fewer than three locations. We assumed movements were influenced by two behavioural states, which we interpreted as representing movements while foraging or travelling. The former state was characterized by less directed movement (i.e. shorter step lengths and larger turn angles) and the latter state by more directed movement (i.e, longer step lengths and smaller turn angles). Conditional on the latent state, we assumed that the distributions of step lengths and turn angles followed gamma and von Mises distributions, respectively. We fit the model using the `fitHMM()` function in the `momentuHMM` package (McClintock & Michelot, 2018). We evaluated model goodness of fit using residual plots constructed using the `plotPR()` function in the same package.

We then created a series of lineups, using the `simData()` function in the `momentuHMM` package (McClintock & Michelot, 2018) to simulate trajectories from the fitted model. We considered lineups depicting qqplots for the distribution of pseudo-residuals for (1) step lengths (Figure S5) and (2) turn angles (Figure 4), (3) simulated and observed trajectories (Figure S6), autocorrelation functions applied to the (4) step-length pseudo-residuals (Figure S7) and (5) turn-angle pseudo-residuals (Figure S8) and (6) circular boxplots showing the distribution of turn angles for different quantiles of the step-length distribution (Figure S9).

To provide an example of how one could easily generate independent lineups for each observer, we fit the same model to three subsets of the data. Each subset included data from a unique combination of two individual fishers, with the third fisher's data withheld for cross-validation. Subsequently, we generated three separate lineups depicting distributions of turn angles using only data from the withheld individual (Figures S10–S12). These lineups could then

be shown to three independent observers and the *p*-value calculated as outlined in Section 2.

4.2 | Results

The `plotPR()` function provides three pseudo-residual plots associated with each of the state-dependent response variables (in this case, observed step lengths and turn angles; Figure 3). These include time series plots of the pseudo-residuals, qqplots comparing their distribution to that of a standard normal distribution, and an autocorrelation plot to evaluate whether the pseudo-residuals are temporally autocorrelated. Confidence bands are automatically included in the qqplots and the autocorrelation plots to help judge whether the pseudo-residuals are independent and consistent with a standard normal distribution.

The Markovian assumption applied to latent behavioural states of the HMM will induce temporal autocorrelation in the step lengths and turn angles. If the Markovian assumption is sufficient for capturing the autocorrelation in the data, then we would expect the pseudo-residuals to be statistically independent. The autocorrelation in the turn angles appears to be well modelled but the pseudo-residuals for the step lengths still exhibit autocorrelation at a few lags, especially at lag 1 (Figure 3). This autocorrelation in the pseudo-residuals might be indicative of a higher-order Markov process. The qqplots of the pseudo-residuals largely fall within the confidence bands, though there are some fairly large pseuoresiduals for the turn angles and an outlier that falls outside of the confidence band. We suspect many users would struggle to interpret whether this outlier is a large or minor issue. Although confidence bands can help with avoiding subjectivity when evaluating residual plots, an advantage of presenting plots in a lineup is that one can see the individual variation that contributes to these bands, and thus, have a better feel for the variation expected across different realizations of the data. Although we suspect many observers would pick the panel holding the real data when presented with the lineup of qqplots of turn-angle

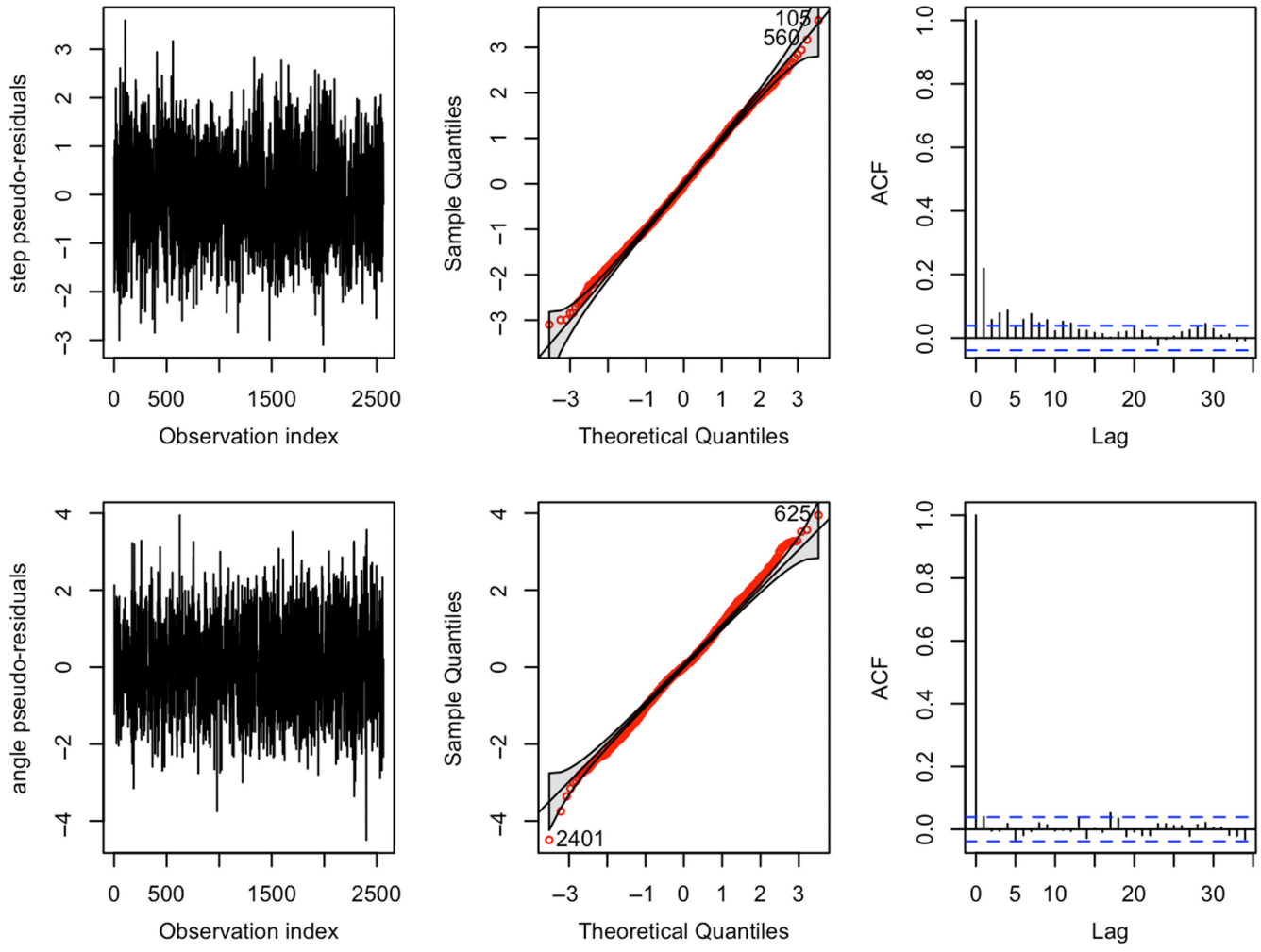

FIGURE 3 Psuedo-residual plots for a two-state hidden Markov model fit to the fisher data from LaPoint et al. (2013a, 2013b). Plots were created using the `plotPR()` function in the `momentuHmm` package (McClintock & Michelot, 2018). The upper and lower rows correspond to step-length and turn-angle pseudo-residuals, respectively. The columns (left, middle, right) depict the time series of the pseudo-residuals, compare their distribution to that of a normal distribution using a quantile-quantile plot, and evaluate whether they are temporally autocorrelated.

pseudo-residuals (Figure 4), there is a second panel that also includes several pseudo-residuals falling outside of the confidence bands.

By constructing the same set of plots for both HMMs and ISSAs, we can compare their relative goodness of fit. For example, in Figure 5, we contrast simulated data from our HMM and our ISSA to evaluate their ability to capture the observed cross-correlation between fisher step lengths and turn angles. The coloured boxes in the the figure capture the middle 50% of the distribution of turn angles for different quantiles of step length, with the outermost ring corresponding to the largest step lengths. If we look at the observed data (panel at the top of Figure 5), we see that turn angles are centred on 0 for these largest step lengths but are otherwise centered on $\pm\pi$. Thus, the fisher tend to either take long and directed steps (large step lengths, turn angles near 0) or shorter and more circuitous steps (smaller step lengths, turn angles frequently near $\pm\pi$). The HMM comes closer than the ISSA at replicating this feature of the data. The ISSA assumes step lengths and turn angles are independent, and thus, there is no consistent

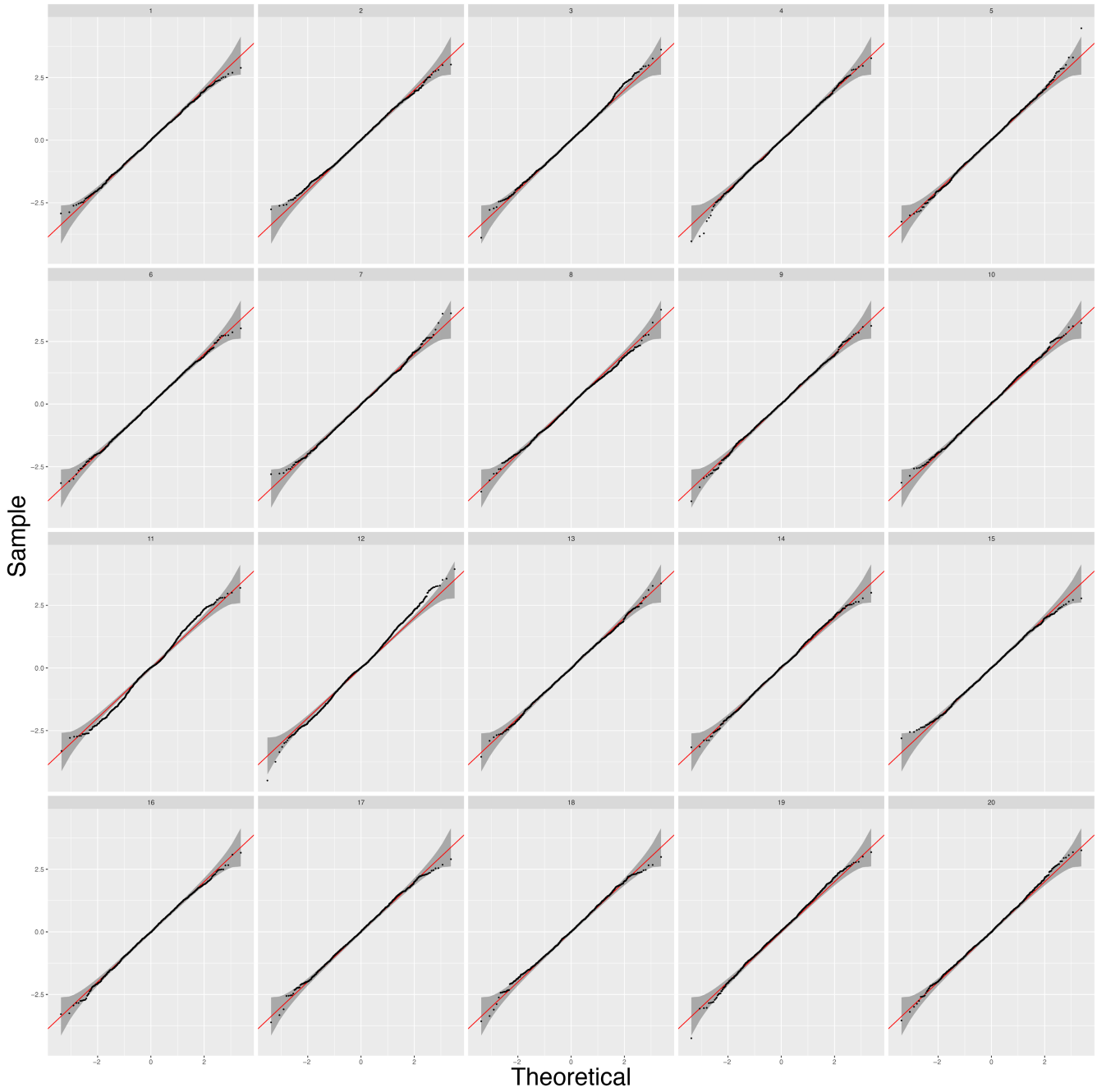

FIGURE 4 Quantile-quantile plots of turn-angle pseudo-residuals from a two-state hidden Markov model (HMMs) fit to the fisher data from LaPoint et al. (2013a, 2013b) and to 19 simulated data sets. The 19 simulated data sets were created using the `simData()` function in the `momentuHMM` package (McClintock & Michelot, 2018) using parameters estimated from fitting the HMM to the fisher data from LaPoint et al. (2013a, 2013b).

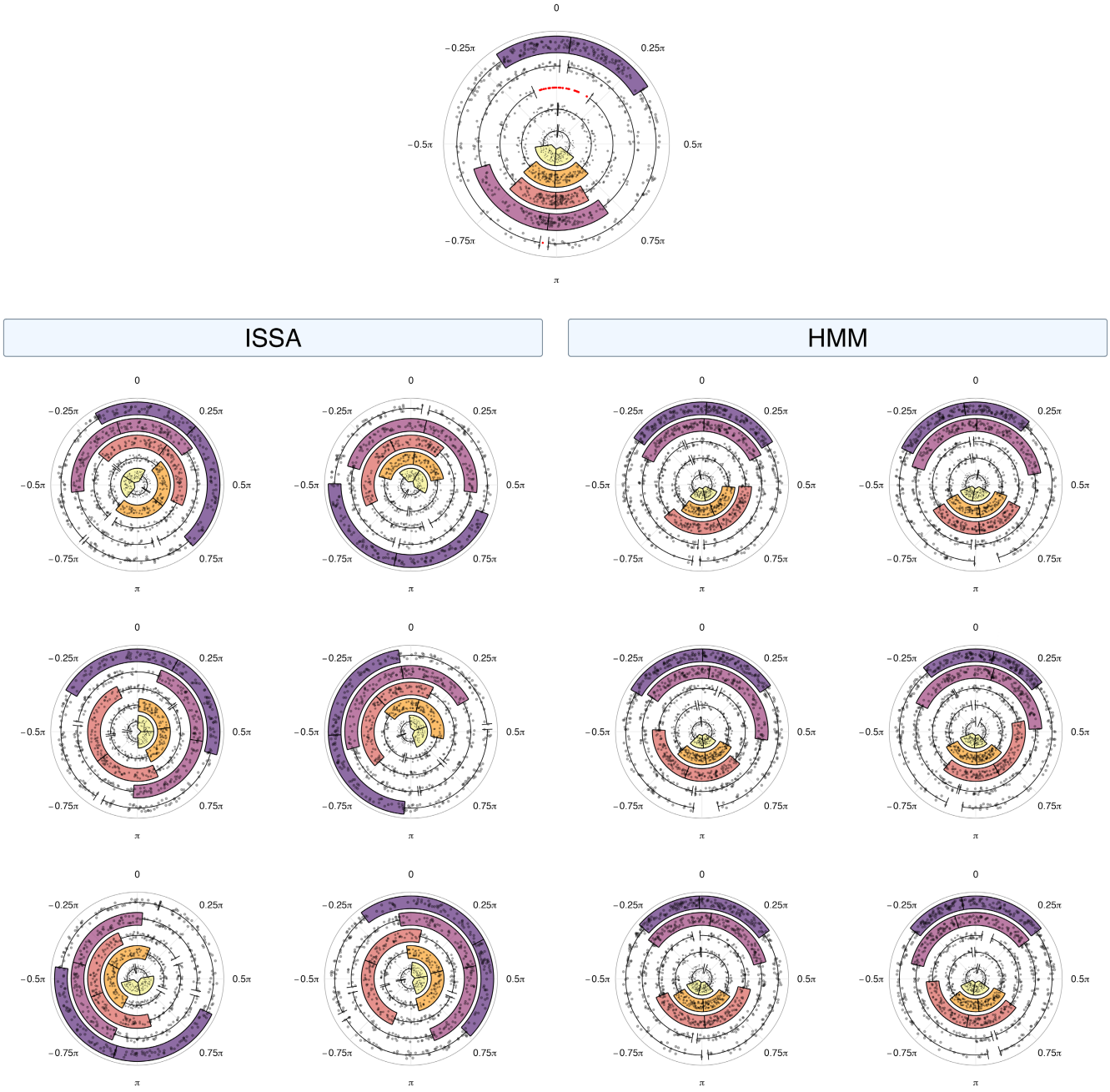

FIGURE 5 Circular boxplots showing the distribution of turn angles for different quantiles of the step-length distribution. The outermost ring (in purple) corresponds to the largest step lengths and the innermost ring to the smallest step lengths. Coloured boxes correspond to the middle 50% of the distribution of turn angles for each quantile of the step-length distribution. The top panel corresponds to the observed movement data of an individual not used to parameterize the model. The left two columns contain plots for data simulated from the fitted ISSA model. The right two columns contain plots for data simulated from the two-state hidden Markov model. Both models were fit to the fisher data from LaPoint et al. (2013a, 2013b).

pattern in the circular boxplots for the data simulated by the ISSA (left two columns of Figure 5). By contrast, the latent states of the HMM induce cross-correlation between step lengths and turn angles. For the HMM (right two columns of Figure 5), the *two* outermost rings (corresponding again to the largest step lengths) tend to be centred on turn angles near 0, whereas the innermost rings are centred on turn angles near $\pm\pi$.

5 | DISCUSSION

We have shown how simulations from animal movement models can be used to form lineups for evaluating model goodness of fit. We focused on integrated step-selection analyses and hidden Markov models due to their widespread use and utility. ISSAs result in a *spatially-explicit*, individual-based model of animal movement

capable of capturing a wide range of mechanisms, including individual responses to local environmental features such as roads and seismic lines (Dickie et al., 2020; Prokopenko et al., 2017; Scrafford et al., 2018), interactions with conspecifics (Schlägel et al., 2019), and familiarity with previously visited locations and migration routes (Kim et al., 2023; Merkle et al., 2019; Oliveira-Santos et al., 2016). As such, ISSAs offer a powerful approach for parameterizing individual-based models of animal movement that can be simulated to model connectivity (Hofmann et al., 2023) or to predict how animals will respond to land use or climate change (Potts et al., 2022; Signer et al., 2017). HMMs, by contrast, rely on latent (unobserved) behavioural states to model temporal variability in animal movement characteristics (e.g. step lengths and turn angles) (Langrock et al., 2012). Applications of HMMs range from cataloguing harbour seal activity budgets (McClintock et al., 2013) to reconstructing bird trajectories based on pressure and wind data (Nussbaumer et al., 2023).

5.1 | Comparison to currently available tools for model evaluation

Despite the widespread use of ISSAs and HMMs in ecology, there are relatively few demonstrations of methods for assessing their goodness of fit, particularly for ISSAs. Fieberg et al. (2018) suggested creating used-habitat-calibration plots based on one-step-ahead predictions, but this approach has some limitations. For example, we observed limited variation in the predicted distribution of environmental covariates (Figure 2a–c). We suspect this will be a common feature associated with one-step-ahead predictions since most environmental variables will be spatially autocorrelated and individuals will have limited movement capacity within a single time step. This limitation can be easily addressed using multi-step simulations, with the caveat that they can be computationally costly to perform. Thus, rather than generate simulation envelopes from a small number of multi-step trajectories, users may instead choose to overlay distributional summaries for each simulation on the same plot.

When considering simulation envelopes and used-habitat-calibration plots for continuous predictors, it is also important to recognize that the type I error will be much greater than α if users reject the null hypothesis whenever the real data summary falls outside the $(1-\alpha)$% simulation envelope (Baddeley et al., 2014; Loosmore & Ford, 2006). To control the type I error, we could instead summarize each density curve using a single summary statistic (e.g. the maximum deviation from the average of all trajectories), from which we could derive an appropriate null distribution. We would then compute the same summary statistic for the observed data and compare it to this null distribution in order to calculate a p-value for the goodness-of-fit test (see e.g. Baddeley et al., 2014).

Another challenge associated with goodness-of-fit tests in general is that to implement them, we often have to first estimate parameters associated with our assumed model. Then, when we simulate data from our fitted model, we are effectively testing whether the data are consistent with the model, as parameterized using the data. This tends to result in conservative goodness-of-fit tests (Baddeley et al., 2014), meaning we are less likely to reject the null hypothesis than if the test were formulated more generally (not conditional on the estimated parameters). To incorporate parameter uncertainty, one could sample a new set of model parameters from a multivariate normal distribution (approximating the asymptotic sampling distribution of the parameters when using maximum likelihood) for each simulated trajectory (see e.g. section 15.4.5 of Fieberg, 2024); this approach offers a frequentist analog to Bayesian posterior predictive distributions. The current implementation of used-habitat-calibration plots in the `amt` package uses this approach to incorporate uncertainty in the habitat-selection parameters when generating one-step-ahead predictions, but it assumes the tentative movement parameters are correct and known. The approach could easily be adapted to use the updated movement kernel and to incorporate uncertainty in the movement parameters.

Methods for assessing fit of HMMs are better developed and largely rely on diagnostic plots created using pseudo-residuals. Judging whether residual plots are problematic enough to warrant concern can be challenging, however, especially for less experienced analysts (Li et al., 2023). Confidence bands can help with avoiding subjectivity. Lineups can also be used to conduct a formal goodness-of-fit test, and they can help less experienced users calibrate their expectations by illustrating the inherent variability expected across replicated data sets (Loy et al., 2016). Although we focused on applications of the lineup protocol for goodness-of-fit testing of animal movement models, the approach can be used much more broadly (see e.g. applications in Buja et al., 2009; Wickham et al., 2010).

We demonstrated, via the HMM application, how lineups can be replicated and shown to multiple users. In the past, some researchers have used Amazon Mechanical Turk for this purpose (Hofmann et al., 2012; Majumder et al., 2014), though we suggest that participatory science platforms, such as Zooinverse (Simpson et al., 2014), offer an appealing alternative. With sufficient observers, lineups have been shown to be competitive with traditional methods for testing hypotheses with linear models (i.e. they have been shown to have similar statistical power) (Majumder et al., 2011, 2013). The advantage of the lineup approach, however, is in its flexibility—that is *any* visual summary of the data can be turned into a test statistic. If the same lineup is shown repeatedly to different observers, then a binomial sampling distribution will no longer be appropriate for calculating p-values, but model-based inferential frameworks are available (VanderPlas et al., 2021).

5.2 | How to choose a visualization

Collectively, the lineups shown here and in the Appendix demonstrate the flexibility of the approach—one can choose *any* graphical summary of the data to test goodness of fit. This begs the question, how should a user go about choosing a particular visual summary when evaluating model fit? One option is to focus on particular features we hope the model will replicate or assumptions about which we are particularly concerned. For example, if one is concerned about cross-correlations between step lengths and turn angles, one can inspect scatterplots

(Figure S3) or consider the circular boxplots of turn angles for different quantiles of the step-length distribution (Figure 5). This highlights an important point, multiple visualizations can be constructed to graphically assess the same assumption, and users are free to choose the visualization that they prefer. To inform this choice, the power of tests using different visualizations can be compared using lineups from simulated data sets that increasingly deviate from the null model (Li et al., 2023; Majumder et al., 2011, 2013).

For ISSAs, there has been relatively little attention given to choosing an appropriate movement kernel, and we suspect the defaults used by many users (e.g. gamma and von Mises distributions for step lengths and turn angles) may provide a poor fit to many real data sets. The typical approach of using a tentative movement kernel to generate random steps when conducting ISSAs also makes it challenging to compare models with different movement kernels; this problem can be solved by using other numerical methods for estimating parameters (Michelot et al., 2024). In our application, the distribution of turn angles was bimodal with peaks near 0 and $\pm\pi$, a feature that was not captured in data simulated by our step-selection model (Figure S2). In addition, our model failed to capture the observed cross-correlation in the step lengths and turn angles (Figures S3 and S4). Notable lack of fit can serve to further motivate new statistical methods that address these concerns. For example, Hodel and Fieberg (2022) developed new methods based on copulae to address concerns regarding cross-correlation between step lengths and turn angles.

An important consideration when evaluating models, particularly those used in conservation and management, is whether they are capable of generating accurate predictions across space or time, referred to as *model transferability* (Helmstetter et al., 2021; Rousseau & Betts, 2022; Yates et al., 2018). By treating individuals as independent sample units for cross-validation, we can test how well models fit to individuals from a restricted subset of environments perform when evaluated using individuals living elsewhere. Thus, lineups, and particularly those that visualize associations between movements and environmental variables (e.g. Figure 1; Figure S1), seem particularly appealing for evaluating model transferability (Fieberg et al., 2018). It is tempting to also consider how lineups might facilitate iterative approaches to model building (sensu Potts et al., 2022), whereby features of the data that are poorly replicated can serve to motivate changes in model structure. One has to be careful with this approach, however, since data-driven decisions can easily result in overfit models that perform poorly when applied to new situations (Fieberg & Johnson, 2015; Harrell, 2001).

5.3 | ISSAs versus HMMs and other models of animal movement

In our application of lineups to the fisher data, we found that the ISSA was capable of generating simulated trajectories that looked like the real trajectory for the 'out-of-sample' fisher (Figure 1). However, some of the simulated trajectories looked quite different from the observed trajectory, and the model failed to reproduce the observed peak at $\pm\pi$ in the distribution of turn angles (Figure S2) or the correlation between observed step lengths and turn angles (Figure 5; Figures S3 and S4). The HMM was able to better capture these features of the data by modelling movements as a function of a latent behavioural state taking on one of two discrete values. These states allow for a mixture of movement types (long and directed vs. short and circuitous), which in turn induces correlation between step lengths and turn angles (Hodel & Fieberg, 2022). Building on these two approaches, one might next consider fitting a state-switching step-selection function (Klappstein et al., 2023; Pohle et al., 2024), though the software for fitting and simulating from this class of model is less developed. Other alternatives could also be explored, such as the multistate Langevin diffusion model developed by McClintock and Lander (2024), which can be simulated using functions in the `momentuHMM` package. In summary, we see lineups as a valuable tool that can be used to evaluate goodness of fit of a wide range of models, and thus fill an important gap when it comes to our ability to evaluate ISSAs, HMMs, and other models of animal movement.

AUTHOR CONTRIBUTIONS

John Fieberg conceived of the original idea for the paper, wrote code for implementing the ISSA with help from Johannes Signer and Smith Freeman, and wrote the first draft of the paper. Smith Freeman analysed the data using hidden Markov models and validated, documented and created functions and vignettes for implementing the lineups. Johannes Signer implemented the used-habitat-calibration plots for the ISSA and helped with writing code for simulating movements in the ISSA. All authors helped revise the paper and gave final approval for submission.

ACKNOWLEDGEMENTS

We thank the Associate Editor and two reviewers, whose comments helped to improve the paper.

FUNDING STATEMENT

JF was supported by National Aeronautics and Space Administration award 80NSSC21K1182 and received partial salary support from the Minnesota Agricultural Experimental Station.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflicts of interest with respect to this paper.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14336.

DATA AVAILABILITY STATEMENT

We have archived the data and code needed to reproduce all analyses in the paper in the Data Repository of the University of Minnesota (Fieberg et al., 2024), https://doi.org/10.13020/3m0d-mp29.

ORCID

John Fieberg https://orcid.org/0000-0002-3180-7021

REFERENCES

Avgar, T., Potts, J. R., Lewis, M. A., & Boyce, M. S. (2016). Integrated step selection analysis: Bridging the gap between resource selection and animal movement. *Methods in Ecology and Evolution*, *7*(5), 619–630.

Baddeley, A., Diggle, P. J., Hardegen, A., Lawrence, T., Milne, R. K., & Nair, G. (2014). On tests of spatial pattern based on simulation envelopes. *Ecological Monographs*, *84*(3), 477–489.

Brown, D. D., LaPoint, S., Kays, R., Heidrich, W., Kümmeth, F., & Wikelski, M. (2012). Accelerometer-informed GPS telemetry: Reducing the trade-off between resolution and longevity. *Wildlife Society Bulletin*, *36*(1), 139–146. https://doi.org/10.1002/wsb.111

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. F., & Wickham, H. (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *367*, 4361–4383.

Center for International Earth Science Information Network (CIESIN) Columbia University, & CIAT, Centro Internacional de Agricultura Tropical. (2005). *Gridded population of the world, version 3 (GPWv3): Population density grid*. NASA Socioeconomic Data; Applications Center (SEDAC). https://doi.org/10.7927/H4XK8CG2

Chatterjee, N., Wolfson, D., Kim, D., Velez, J., Freeman, S., Bacheler, N. M., Shertzer, K., Taylor, J. C., & Fieberg, J. (2024). Modeling individual variability in habitat selection and movement using integrated step-selection analyses. *Methods in Ecology and Evolution*. https://doi.org/10.1111/2041-210X.14321

Conn, P. B., Johnson, D. S., Williams, P. J., Melin, S. R., & Hooten, M. B. (2018). A guide to bayesian model checking for ecologists. *Ecological Monographs*, *88*(4), 526–542.

Defourny, P., Schouten, L., Bartalev, S., Bontemps, S., Cacetta, P., De Wit, A., Di Bella, C., Gerard, B., Giri, C., Gond, V., Hazeu, G., Heinimann, A., Herold, M., Knoops, J., Jaffrain, G., Latifovic, R., Lin, H., Mayaux, P., Mücher, S., … Arino, O. (2009). *Accuracy assessment of a 300 m global land cover map: The GlobCover Experience*. International Center for Remote Sensing of Environment (ICRSE).

Dickie, M., McNay, S. R., Sutherland, G. D., Cody, M., & Avgar, T. (2020). Corridors or risk? Movement along, and use of, linear features varies predictably among large mammal predator and prey species. *Journal of Animal Ecology*, *89*(2), 623–634.

Fieberg, J. (2024). *Statistics for ecologists: A frequentist and bayesian treatment of modern regression models*. University of Minnesota Libraries Publishing. https://hdl.handle.net/11299/260227

Fieberg, J., Forester, J. D., Street, G. M., Johnson, D. H., ArchMiller, A. A., & Matthiopoulos, J. (2018). Used-habitat calibration plots: A new procedure for validating species distribution, resource selection, and step-selection models. *Ecography*, *41*(5), 737–752.

Fieberg, J., Freeman, S., & Signer, J. (2024). R code associated with evaluating goodness-of-fit of animal movement models using lineups. Data Repository for the University of Minnesota. https://doi.org/10.13020/3m0d-mp29

Fieberg, J., & Johnson, D. H. (2015). MMI: Multimodel inference or models with management implications? *The Journal of Wildlife Management*, *79*(5), 708–718.

Fieberg, J., Signer, J., Smith, B., & Avgar, T. (2021). A 'how to' guide for interpreting parameters in habitat-selection analyses. *Journal of Animal Ecology*, *90*(5), 1027–1043.

Harrell, F. E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis* (Vol. *608*). Springer.

Helmstetter, N. A., Conway, C. J., Stevens, B. S., & Goldberg, A. R. (2021). Balancing transferability and complexity of species distribution models for rare species conservation. *Diversity and Distributions*, *27*(1), 95–108.

Hodel, F. H., & Fieberg, J. (2021). Cylcop: An r package for circular-linear copulae with angular symmetry. *bioRxiv*, https://doi.org/10.1101/2021.07.14.452253

Hodel, F. H., & Fieberg, J. (2022). Circular-linear copulae for animal movement data. *Methods in Ecology and Evolution*, *13*, 1001–1013. https://doi.org/10.1111/2041-210X.13821

Hofmann, D. D., Cozzi, G., McNutt, J. W., Ozgul, A., & Behr, D. M. (2023). A three-step approach for assessing landscape connectivity via simulated dispersal: African wild dog case study. *Landscape Ecology*, *38*(4), 981–998.

Hofmann, H., Follett, L., Majumder, M., & Cook, D. (2012). Graphical tests for power comparison of competing designs. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2441–2448.

Hooten, M. B., Johnson, D. S., McClintock, B. T., & Morales, J. M. (2017). *Animal movement: Statistical models for telemetry data*. CRC Press.

Joo, R., Boone, M. E., Clay, T. A., Patrick, S. C., Clusella-Trullas, S., & Basille, M. (2020). Navigating through the r packages for movement. *Journal of Animal Ecology*, *89*(1), 248–267.

Kays, R., Crofoot, M. C., Jetz, W., & Wikelski, M. (2015). Terrestrial animal tracking as an eye on life and planet. *Science*, *348*(6240), aaa2478.

Kim, D., Thompson, P., Wolfson, D., Merkle, J., Oliveira-Santos, L. G. R., Forester, J., Avgar, T., Lewis, M. A., & Fieberg, J. (2023). Identifying signals of memory from observations of animal movements in plato's cave. *bioRxiv*, 2023–08. https://doi.org/10.1101/2023.08.15.553411

Klappstein, N. J., Thomas, L., & Michelot, T. (2023). Flexible hidden markov models for behaviour-dependent habitat selection. *Movement Ecology*, *11*(30), 13.

Langrock, R., King, R., Matthiopoulos, J., Thomas, L., Fortin, D., & Morales, J. M. (2012). Flexible and practical modeling of animal telemetry data: Hidden Markov models and extensions. *Ecology*, *93*(11), 2336–2342.

LaPoint, S., Gallery, P., Wikelski, M., & Kays, R. (2013a). Animal behavior, cost-based corridor models, and real corridors. *Landscape Ecology*, *28*(8), 1615–1630. https://doi.org/10.1007/s10980-013-9910-0

LaPoint, S., Gallery, P., Wikelski, M., & Kays, R. (2013b). Data from: Animal behavior, cost-based corridor models, and real corridors. *Movebank Data Repository*, https://doi.org/10.5441/001/1.2tp2j43g

Li, W., Cook, D., Tanaka, E., & VanderPlas, S. (2023). A plot is worth a thousand tests: Assessing residual diagnostics with the lineup protocol. *arXiv Preprint arXiv:2308.05964*.

Loosmore, N. B., & Ford, E. D. (2006). Statistical inference using the g or k point pattern spatial statistics. *Ecology*, *87*(8), 1925–1931.

Loy, A., Follett, L., & Hofmann, H. (2016). Variations of q–q plots: The power of our eyes! *The American Statistician*, *70*(2), 202–214.

Majumder, M., Hofmann, H., & Cook, D. (2011). *Visual statistical inference for regression parameters*. Technical Report 13, Iowa State University, Department of Statistics.

Majumder, M., Hofmann, H., & Cook, D. (2013). Validation of visual statistical inference, applied to linear models. *Journal of the American Statistical Association*, *108*(503), 942–956.

Majumder, M., Hofmann, H., & Cook, D. (2014). Human factors influencing visual statistical inference. *arXiv Preprint arXiv:1408.1974*.

McClintock, B. T., King, R., Thomas, L., Matthiopoulos, J., McConnell, B. J., & Morales, J. M. (2012). A general discrete-time modeling framework for animal movement using multistate random walks. *Ecological Monographs*, *82*(3), 335–349.

McClintock, B. T., & Lander, M. E. (2024). A multistate l angevin diffusion for inferring behavior-specific habitat selection and utilization distributions. *Ecology*, *105*(1), e4186.

McClintock, B. T., & Michelot, T. (2018). momentuHMM: R package for generalized hidden markov models of animal movement. *Methods in Ecology and Evolution*, *9*(6), 1518–1530.

McClintock, B. T., Russell, D. J. F., Matthiopoulos, J., & King, R. (2013). Combining individual animal movement and ancillary biotelemetry data to investigate population-level activity budgets. *Ecology*, *94*(4), 838–849. https://doi.org/10.1890/12-0954.1

Merkle, J. A., Sawyer, H., Monteith, K. L., Dwinnell, S. P., Fralick, G. L., & Kauffman, M. J. (2019). Spatial memory shapes migration and its benefits: Evidence from a large herbivore. *Ecology Letters*, *22*(11), 1797–1805.

Michelot, T., Klappstein, N. J., Potts, J. R., & Fieberg, J. (2024). Understanding step selection analysis through numerical integration. *Methods in Ecology and Evolution*, *15*(1), 24–35.

Muff, S., Signer, J., & Fieberg, J. (2020). Accounting for individual-specific variation in habitat-selection studies: Efficient estimation of mixed-effects models using bayesian or frequentist computation. *Journal of Animal Ecology*, *89*(1), 80–92.

Nussbaumer, R., Gravey, M., Briedis, M., Liechti, F., & Sheldon, D. (2023). Reconstructing bird trajectories from pressure and wind data using a highly optimized hidden markov model. *Methods in Ecology and Evolution*, *14*(4), 1118–1129. https://doi.org/10.1111/2041-210X.14082

Oliveira-Santos, L. G. R., Forester, J. D., Piovezan, U., Tomas, W. M., & Fernandez, F. A. (2016). Incorporating animal spatial memory in step selection functions. *Journal of Animal Ecology*, *85*(2), 516–524.

Pedersen, T. L. (2020). *Patchwork: The composer of plots*. https://CRAN.R-project.org/package=patchwork

Pohle, J., Signer, J., Eccard, J. A., Dammhahn, M., & Schlägel, U. E. (2024). How to account for behavioral states in step-selection analysis: A model comparison. *PeerJ*, *12*, e16509.

Potts, J. R., Börger, L., Strickland, B. K., & Street, G. M. (2022). Assessing the predictive power of step selection functions: How social and environmental interactions affect animal space use. *Methods in Ecology and Evolution*, *13*(8), 1805–1818.

Prokopenko, C. M., Boyce, M. S., & Avgar, T. (2017). Characterizing wildlife behavioural responses to roads using integrated step selection analysis. *Journal of Applied Ecology*, *54*(2), 470–479.

Rousseau, J. S., & Betts, M. G. (2022). Factors influencing transferability in species distribution models. *Ecography*, *2022*(7), e06060.

Schlägel, U. E., Signer, J., Herde, A., Eden, S., Jeltsch, F., Eccard, J. A., & Dammhahn, M. (2019). Estimating interactions between individuals from concurrent animal movements. *Methods in Ecology and Evolution*, *10*(8), 1234–1245.

Scrafford, M. A., Avgar, T., Heeres, R., & Boyce, M. S. (2018). Roads elicit negative movement and habitat-selection responses by wolverines (*Gulo gulo luscus*). *Behavioral Ecology*, *29*(3), 534–542.

Signer, J., Fieberg, J., & Avgar, T. (2017). Estimating utilization distributions from fitted step-selection functions. *Ecosphere*, *8*(4), e01771.

Signer, J., Fieberg, J., & Avgar, T. (2019). Animal movement tools (amt): R package for managing tracking data and conducting habitat selection analyses. *Ecology and Evolution*, *9*(2), 880–890.

Signer, J., Fieberg, J., Reineking, B., Schlaegel, U., Smith, B. J., Balkenhol, N., & Avgar, T. (2023). Simulating animal space use from fitted integrated step-selection functions (iSSF). *bioRxiv*, 2023–08.

Simpson, R., Page, K. R., & De Roure, D. (2014). Zooniverse: Observing the world's largest citizen science platform. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 1049–1054). https://doi.org/10.1145/2567948.2579215

U.S./Japan ASTER Science Team. (2009). *ASTER global digital elevation model data set*. NASA EOSDIS Land Processes DAAC. https://lpdaac.usgs.gov/dataset_discovery/aster/aster_products_table/astgtm_v002, https://doi.org/10.5067/ASTER/ASTGTM.002

VanderPlas, S., Röttger, C., Cook, D., & Hofmann, H. (2021). Statistical significance calculations for scenarios in visual inference. *Stat*, *10*(1), e337.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://ggplot2.tidyverse.org

Wickham, H., Cook, D., Hofmann, H., & Buja, A. (2010). Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 973–979.

Wilke, C. O., & Wiernik, B. M. (2022). *Ggtext: Improved text rendering support for 'ggplot2'*. https://CRAN.R-project.org/package=ggtext

Yates, K. L., Bouchet, P. J., Caley, M. J., Mengersen, K., Randin, C. F., Parnell, S., Fielding, A. H., Bamford, A. J., Ban, S., Barbosa, A. M., Dormann, C. F., Elith, J., Embling, C. B., Ervin, G. N., Fisher, R., Gould, S., Graf, R. F., Gregr, E. J., Halpin, P. N., … Sequeira, A. M. M. (2018). Outstanding challenges in the transferability of ecological models. *Trends in Ecology & Evolution*, *33*(10), 790–802.

SUPPORTING INFORMATION

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

Figure S1: Distribution of elevation measured at the end of each movement step.

Figure S2: Distribution of turn angles.

Figure S3: Scatterplot of step lengths (*x*-axis) and turn angles (*y*-axis).

Figure S4: Distribution of turn angles for different quantiles of step length.

Figure S5: Quantile-quantile plots of step-length pseudo-residuals from hidden Markov models fit to the fisher data and to 19 simulated data sets.

Figure S6: Simulated trajectories.

Figure S7: ACF plots of the psuedo residuals for step lengths.

Figure S8: ACF plots of the psuedo residuals for turn angles.

Figure S9: Distribution of turn angles for different quantiles of step length.

Figure S10: Distribution of turn angles.

Figure S11: Distribution of turn angles.

Figure S12: Distribution of turn angles.

How to cite this article: Fieberg, J., Freeman, S., & Signer, J. (2024). Using lineups to evaluate goodness of fit of animal movement models. *Methods in Ecology and Evolution*, *00*, 1–12. https://doi.org/10.1111/2041-210X.14336