A Framework for Studying Animal Behavior Using Deep Learning

Marcel Gietzmann-Sanders, Michael Courtney, Andrew Seitz, Curry Cunningham

University of Alaska Fairbanks

This study explores using deep learning to build Individual-Based Models (IBMs). Using Chinook salmon (*Oncorhynchus tshawytscha*) PSAT data as a case study, we introduce a methodology that leverages deep learning to transform the construction of IBMs into a systematic tool for investigating behaviors and their covariates. By employing a "log-odds model" to reduce dimensionality, this approach positions modeling as a central driver of discovery, enabling efficient hypothesis testing and iterative refinement.

Introduction

Modeling organism behavior can be thought of as equivalent to modeling choices and selections. For example if modeling animal movement, we can imagine each timestep as a decision d_i where the animal is selecting from a series of grid cells in its vicinity. Each such grid cell would represent a choice $c_{i,j}$. For each of these cells we may have a series of features $\vec{v}_{i,j}$ such as environmental covariates, descriptors such as distance or direction from the animal, and/or features of the animal itself or its state. Then we could pose our model as the function $F(\vec{v}_{i,j})$ which gives the probability of selecting $c_{i,j}$ given the information at hand. That is, we model the animal's behavior as a discrete Markov process.

The fact that this process produces probabilities per choice as opposed to explicit predictions of behavior is advantageous for two reasons:

1. With a deterministic model that predicts a specific trajectory you require a high level of $\vec{v}_{i,j}$ to use the model beyond a small number of steps as errors can accumulate. Here because we are predicting probabilities across all trajectories we can see what $\vec{v}_{i,j}$ elucidates and what it doesn't.

2. Because a specific likelihood is provided for each choice and trajectory we can identify specific decisions, individual organisms, places, or periods of time in which the model gives low likelihoods relative to the rest of the data making the model a tool for exploratory data analysis as well.

The fact that we have posed it in terms of distinct choices means that this is precisely a probabilistic machine learning problem (Oliver Durr, 2020) meaning that we can take advantage of machine learning to learn $F(\vec{v}_{i,j})$. This carries with it a few advantages as well.

1. Machine learning models, especially deep learning models, allow for very flexible pattern matching.

2. Rather than having to rebuild a model for each addition or subtraction to $\vec{v}_{i,j}$ we can reuse the same tool set to build and validate models regardless of $\vec{v}_{i,j}$.

This combination presents a kind of virtuous cycle in that the modeling process can be used to provide the very exploratory data analysis that can point to improvements in the model itself thus allowing model development to supercharge itself.

There is significant precedent here as Hidden Markov Models (HMMs) and machine learning have both seen extensive use in movement modeling (Dhanushi A. Wijeyakulasuriya, 2020). However while each of these in turn bring some of the advantages outlined above they do not typically combine them.

Hidden Markov Models, by definition, do provide a conditioned probability distribution but the examples from the literature require explicit assumptions about how the data and behavior interact and therefore aren't able to take advantage of the flexibility and automation provided by machine learning (Dhanushi A. Wijeyakulasuriya, 2020). Therefore they provide they can technically provide the first set of advantages but not the second.

The applications of machine learning in the space do the opposite. They allow for the flexible and automated training of models, but those models typically make explicit predictions on movement, such as regressions on step length and angle thereby providing the latter set of advantages but not the former (Dhanushi A. Wijeyakulasuriya, 2020)(Daniel Clark and Clark, 2021)(Daniel Einarson and Sennersten, 2024).

The aim of this paper is to provide a guide on how to use probabilistic machine learning to model animal behavior as well as an example application

to movement data in order to demonstrate ways in which the methodology can be used as an exploratory data analysis tool as well.

Applying Probabilistic Machine Learning to Behavior

Theory

Standard probabilistic deep learning networks are typically framed as a classification problem, using categorical cross-entropy as the loss function (Oliver Durr, 2020). Each output neuron represents a potential choice, with the model predicting the probability of each choice being correct based on this loss formulation. For these choices, we provide the network with features encapsulating the relevant information. Training is then comprised of providing a series such decisions.

However, this formulation introduces a critical challenge: if there are N features per choice and M potential choices, the overall dimensionality of the input space becomes $N \cdot M$. Adding even a single feature increases the dimensionality by M not just 1.

This growth poses a significant challenge due to the "curse of dimensionality", where the amount of data required to effectively train models can grow exponentially with the dimensionality of the input space (Verleysen and François, 2005).

Log-Odds Modeling

To address this issue we could take advantage of the order invariance of choices in the traditional probabilistic problem framing. Specifically, the order in which choices are presented to the model should not matter. For instance, whether a particular choice appears in the first or the thirteenth position should have no impact on the model's operation. This property allows for data augmentation by reordering choices.

In essence, for each training example, $M!$ (factorial of the number of choices) augmented examples can be created.

The issue with this approach is that as your augmented data size grows to match the needs of the greater dimensionality, so too does the time required for training. So while the problem remains theoretically possible, the potential exponential uptick in time complexity poses a significant practical issue.

Instead we propose an adjustment to the standard framing of probabilistic machine learning. Instead of predicting the probabilities directly, we predict the log-odds ϕ_m for each choice and calculate the probability p_m using the softmax function:

$$p_m = \frac{e^{\phi_m}}{\sum_{m=1}^{M} e^{\phi_m}}$$

This approach reduces the feature space dimensionality to N and effectively increases the number of training examples by a factor of M.

We can implement this log-odds model using standard probabilistic deep learning techniques by replicating the "log-odds model" weights across all M choices. The outputs are fed into a softmax layer with M units, where the layer's weights are set to the identity matrix and biases are set to zero. Using categorical cross-entropy as the loss function ensures compatibility with standard probabilistic deep learning while enabling us to train the log-odds weights and significantly reduce the problem's dimensionality.

Contrast Sampling

A practical issue with our log-odds framing is that as M grows large most instances of the internal log-odds model would ideally report very low log-odds, resulting in low probabilities. Ideally, only one choice should produce $p_m = 1$. This is analogous to a class imbalance problem, where the model becomes prone to predicting the most common class.

To address this, we balance the training data. Instead of presenting the model with full decisions containing all M choices, we create training pairs, or contrasts, where each pair consists of one selected choice and one unselected choice. This approach is valid because the log-odds model focuses on the relative likelihood of choices, making the number of choices considered at any one time irrelevant.

The primary risk in using contrasts is introducing bias by disproportionately sampling certain combinations of choices. To mitigate this, we randomly sample pairs from each decision, ensure an equal number of contrasts per decision, and an equal number of decisions per individual. This preserves the balance across the training data and avoids skewing the model's predictions.

Application to Chinook Salmon Movement Data

Outline

We illustrate this technique using an example path of inquiry, over Chinook movement data, that reflects the selections and considerations involved in behavioral modeling.

In our initial exploratory data analysis we may notice two patterns of movement illustrated by figure 1. Specifically, we notice that there are individuals who seem to "randomly walk" about their region and other individuals who move with reasonably clear direction. From this we may wonder whether to what extent these patterns can be explained using covariates related to productivity.

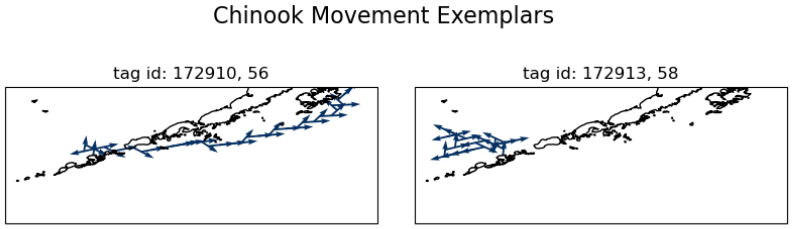

Figure 1: Movements of two sample Chinook salmon off the coast of the Aleutian Islands chain. The individual on the left demonstrates far more directed movement than the individual on the right.

To proceed with this investigation we must first build a null model to use as a basis for comparison, then build a model that can pick up on any latent directionality across our dataset, and finally a model that will allow us to investigate the degree to which productivity can help explain the deviations from this latent directionality. Our log-odds modeling framework will give us both the tools required to build the model efficiently as well as inspect the details of each new covariate's impact.

Data

To demonstrate this methodology in practice, we consider a series of tracks from 111 Chinook salmon (*Oncorhynchus tshawytscha*) caught and monitored between 2013 and 2022 (Michael B Courtney, 2021) (Michael B Courtney, 2019). These tracks were obtained from pop-up satellite archival tags

which collect temperature, light level, and depth information at specified (sub day) intervals. This data is then passed through a proprietary algorithm from Wildlife Computers to determine likely longitude and latitude during each day of of monitoring (WildlifeComputers, 2024).

Environmental data was derived from the Global Ocean Biogeochemistry Hindcast dataset (10.48670/moi-00019) and the Global Ocean Physics Reanalysis (10.48670/moi-00021) from the E.U. Copernicus Marine Service Information. Net primary production (mg/m3/day) and mixed layer thickness (m) were aggregated per Uber h3 resolution 4 cell in the Northern Pacific.

Formulation

The resolution 4 Uber h3 cell containing each salmon location was identified and then, assuming a maximum travel distance of 100km (centeroid to centeroid) all adjacent cells within the 100km were identified as choices (including the currently occupied cell). In general this represented $\sim$ 19 choices per decision with the intention being to predict the probability of moving to any particular cell. Training data was derived by identifying the actual cell moved to.

Features

Movement heading in radians and distance to the centroid of the choice cell were computed and then mixed layer thickness and net primary production were joined to the choices on cell and day.

Distance was normalized to a range of 0-1 by division by 100, while mixed layer thickness and net primary production were both log-scaled and then centered at zero.

Contrast Sampling

After inspecting the distribution of number of choices per salmon and number of choices per decision, we decided on random sampling (with replacement) 200 decisions per individual and 19 choices per decision.

Over a validation/training split of 40, 71 this resulted in 421,800 contrasts of which 269,800 were used in training and the rest in validation.

Note that only 14,200 training examples would've been available to a traditional probabilistic approach representing a large increase in the number of available training examples.

Training

Three sets models were trained, one including only distance, one with both distance and movement heading, and one with all four features. Note that while the feature dimensions of these models are 1, 2, and 4 respectively, given the maximum number of choices per decision seen was 33, the dimensionality of a standard probabilistic model would've been 33, 66, and 132 representing a large reduction in the dimensionality of our feature spaces.

Architectures/hyperparameters for the log-odds component of the model were parametrized in the following ways:

Component	Options
Layers	3, 4
Units per Layer	24, 32
Batch Size	10000
Learning Rate	0.0005

With 5 models trained for each combination. Models were trained in Keras using an Adam optimizer for 100 epochs.

Models for each set of features we selected on the basis of the loss over the validation set of contrasts.

Compute

Models were trained using AWS Batch using Fargate instances of 2 vcpu's and 4 GB of memory. By taking advantage of AWS Batch, models could be all trained in parallel allowing for short (15-30 minute) turn around times.

Results

In training the metric of interest is the contrasts' normalized log probability (C-NLP) - the average log probability per contrast. For an equivalent metric over all of the decisions we also computed the average log probability per decision for each individual and then computed an average over those across individuals (in order to not favor individuals with many decisions). This is the D-NLP reported in the table below.

Model	Train C-NLP	Val C-NLP	Train D-NLP	Val D-NLP
No Model	-0.693	-0.693	-2.944	-2.944
Model 1	-0.172	-0.154	-1.336	-1.223
Model 2	-0.156	-0.150	-1.281	-1.200
Model 3	-0.147	-0.146	-1.248	-1.180

"No Model" assumes all decisions are equally likely, Model 1 is the distance only model, Model 2 adds the movement heading, and Model 3 adds the net productivity and mixed layer thickness features.

Figure 2: Log likelihood increase (unnormalized) per individual of each model change.

Figure 3: Log likelihood increase (unnormalized) per individual of each model change in cases where the selection involved no movement.

Figure 4: Log likelihood increase (unnormalized) per individual of each model change in cases where the selection involved movement.

Figure 2 shows the unnormalized changes in log likelihood for each individual colored by the region in which they were initially tagged as we go

from one model to the next. Figures 3 and 4 break this down for decisions that resulted in no movement and movement respectively. Only individuals in the validation set were considered.

The appendix contains a series of mapped examples (figures 5-10) from each of these regions (besides EBS) that shows how the log likelihood per decision (when moving) changed in moving from model 1 to model 2 and then from model 2 to model 3. As in the above, only individuals in the validation set were considered.

Finally as much of the behavior seems to be modulated by movement distances (or lack of movement) a summary table of empirical likelihood of movement of a specific distance is given (taken over all individuals).

Distance Bin (km)	Likelihood of Selection	# Training Decisions
No Movement	65.6%	3163
Up to 50km	32.5%	1567
50km to 100km	1.9%	92

Discussion

Model One: Distance as a Key Feature

Given the observations that most decisions involve an individual staying put (within its h3 cell) and rarely do fish travel to an h3 cell further than 50km away, it was decided that a more reasonable baseline model than even odds would be one that included distance traveled as a feature (model 1).

As expected model 1 resulted in a significant improvement in log-likelihood across all individuals in the validation set as compared to a model that gives even odds to each available choice. This model therefore provides the baseline for our subsequent models.

Model Two: Movement Heading

Beyond this underlying distance feature, the interest is in predicting, when fish move, where they will move. Of particular interest was the individuals in the sitka and yakutat groups as these groups had several individuals exhibiting what appeared to be very directed and extensive movement.

Upon studying these individuals it was noted that changes in mixed layer thickness seemed to incite movement and the individuals tended to stay (anecdotally) near higher primary productivity. Therefore a model with net primary productivity and mixed layer thickness was of interest.

However, before moving to this model we wanted a baseline model that included movement heading as a feature so that we could determine what aspects of the movement were predicted simply by common headings as opposed to movement specifically related to productivity. Model 2 is this distance plus movement heading model.

Looking at figure 2 we can see that for several individuals in the yakutat group adding movement heading increased the log likelihood meaningfully as compared to model 1 (our baseline). Indeed figure 4 indicates that most of this increase comes from decisions where the individuals moved. Looking at figure 10 it is clear that the move from model 1 to model 2 is creating a model that's learning a heading south east is more likely than alternative directions with the movements of fish in north or westerly directions actually reducing in log likelihood as compared to the baseline.

This southeasterly pattern is maintained throughout figures 5-10 and indeed looking at 4 there are several groups where the majority of individuals suffer a decrease in log likelihood, in all likelihood due to this common direction not being representative of them.

The sitka group however does see an improvement as well showing that both of these majority south eastern movement groups (sitka and yakutat) can be at least partially explained by a common heading down the coast.

Model Three: Environmental Features

With this naive heading baseline in place we can now investigate the extent to which adding in net primary productivity and mixed layer thickness allows our model to distinguish when different headings are taken by different fish, or in fact reinforce the headings already "selected".

Looking at the step from model 2 to model 3 in figures 5-10 we see now that several of the movements that saw a drop in log likelihood as compared to the model 1 baseline are actually seeing a rise in log-likelihood under this new model indicating that the new features are allowing us to swing the preferred movement heading in a way that is conditioned on the environment.

In addition, especially in the sitka group, we see reinforcement/refinement of the decisions as there is an increase in log likelihood up and beyond what was received from model 2. That is to say that while there is a general tendency in the south easterly direction, the specific cells chosen are in some way correlated to productivity.

Noteable as well is the fact that the move from model 2 to model 3 had a significant positive impact on the predictability of the decisions from the

nanwalek group, much more so than in the step from model 1 to model 2. The kodiak group however remains largely unaffected.

Further Questions

Looking at the cases in figure 2 in which we see significant drops in log likelihood or cases where there is no meaningful change we can see that while there is predictive power in the features chosen thus far - especially for the yakutat, sitka, and nanwalek groups - there are still plenty of individuals that can act as exemplars for searching for new behavioral drivers.

For example in figure 5, even with the addition of the productivity features in model 3 the westerly movements are still poorly predicted and would warrant investigation.

Likewise the unalaska group is poorly explained in general and would represent another case study for further investigation. Perhaps by training a model with the same set of features as model 3 but only over this group we would see a significant improvement in the performance of the model which may lead us to look for demographic or geophysical differences that could explain how two different groups of fish are "using" the same covariates so differently.

All in all though, the models we have allow us to identify and focus upon individuals that so far resist our attempts to explain them.

Conclusion

This investigation demonstrates the power and flexibility of an iterative modeling approach for modeling behavior. By building baseline models and progressively refining them with additional covariates, we identify both predictive patterns and individuals resistant to explanation, which serve as focal points for further hypothesis generation. The efficiency of this process, enabled by quick training times and parallelization, ensures that refinement is limited only by the investigator's ability to propose and test new hypotheses. Crucially, at every stage, the models not only advance understanding but remain actionable for predictions, underscoring their dual utility as tools for both discovery and application.

Code

The tooling used to build these models as well as the means to deploy them using Amazon Web Services has been packaged at:
https://github.com/networkearth/mimic

References

Daniel Clark, David Shaw, A. V. S. W. J. S. and Clark, J. D. (2021). Using machine learning methods to predict the movement trajectories of the using machine learning methods to predict the movement trajectories of the louisiana black bear louisiana black bear. *SMUDataScienceReview*, 5.

Daniel Einarson, Fredrik Frisk, K. K. and Sennersten, C. (2024). A machine learning approach to simulation of mallard movements. *Applied Sciences*, 14.

Dhanushi A. Wijeyakulasuriya, Elizabeth W. Eisenhauer, B. A. S. E. M. H. (2020). Machine learning for modeling animal movement. *PLOS One*.

Michael B Courtney, Mark D Evans, J. F. S. A. H. R. A. C. S. (2019). Behavior and thermal environment of chinook salmon oncorhynchus tshawytscha in the north pacific ocean, elucidated from pop-up satellite archival tags. *Environmental Biology of Fishes*, 102:1039–1055.

Michael B Courtney, Mark Evans, K. R. S. A. C. S. (2021). Understanding the behavior and ecology of chinook salmon (oncorhynchus tshawytscha) on an important feeding ground in the gulf of alaska. *Environmental Biology of Fishes*, 104:357–373.

Oliver Durr, Beate Sick, E. M. (2020). *Probabilistic Deep Learning*. Manning Publications.

Verleysen, M. and François, D. (2005). The curse of dimensionality in data mining and time series prediction. *Computational Intelligence and Bioinspired Systems*, 3512:758–770.

Volker Grimm, S. R. (2005). *Individual-based Modeling and Ecology*. Princeton University Press.

WildlifeComputers (2024). Minipat.

Appendix: Mapped Examples

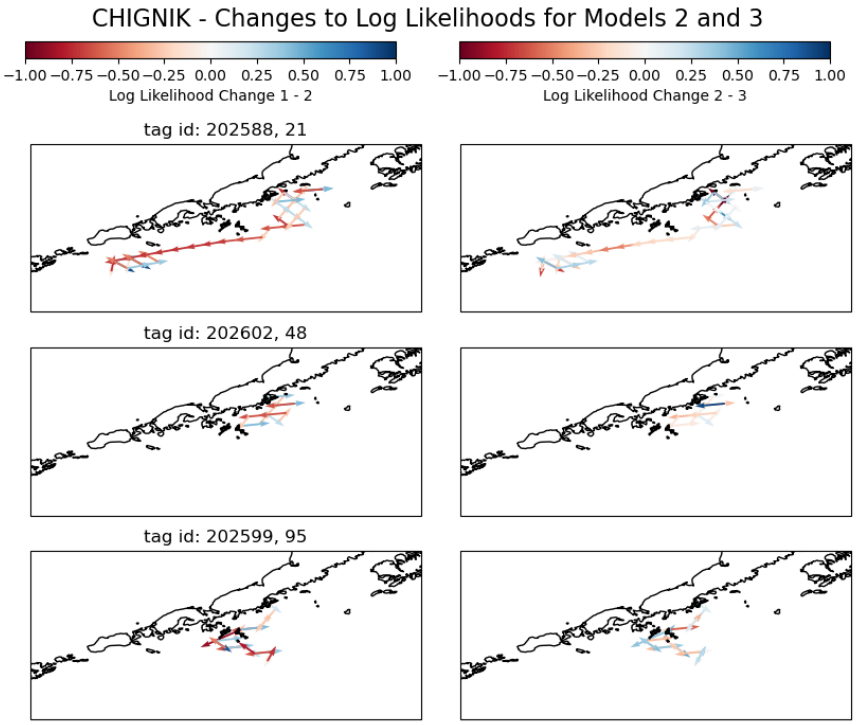

Figure 5: Log likehood changes per decision - Chignik

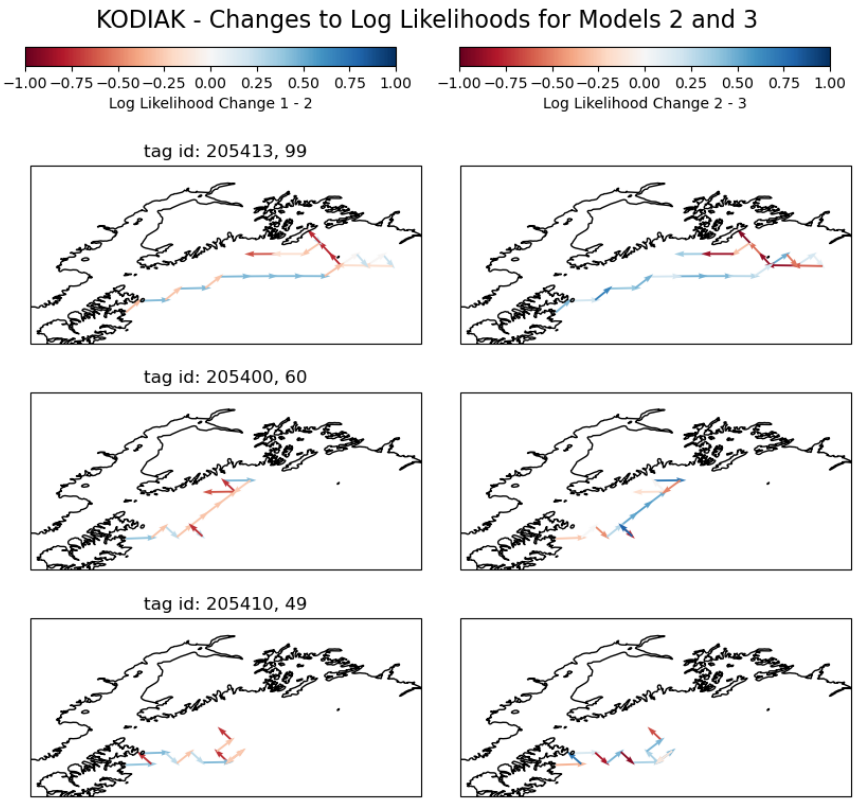

Figure 6: Log likehood changes per decision - Kodiak

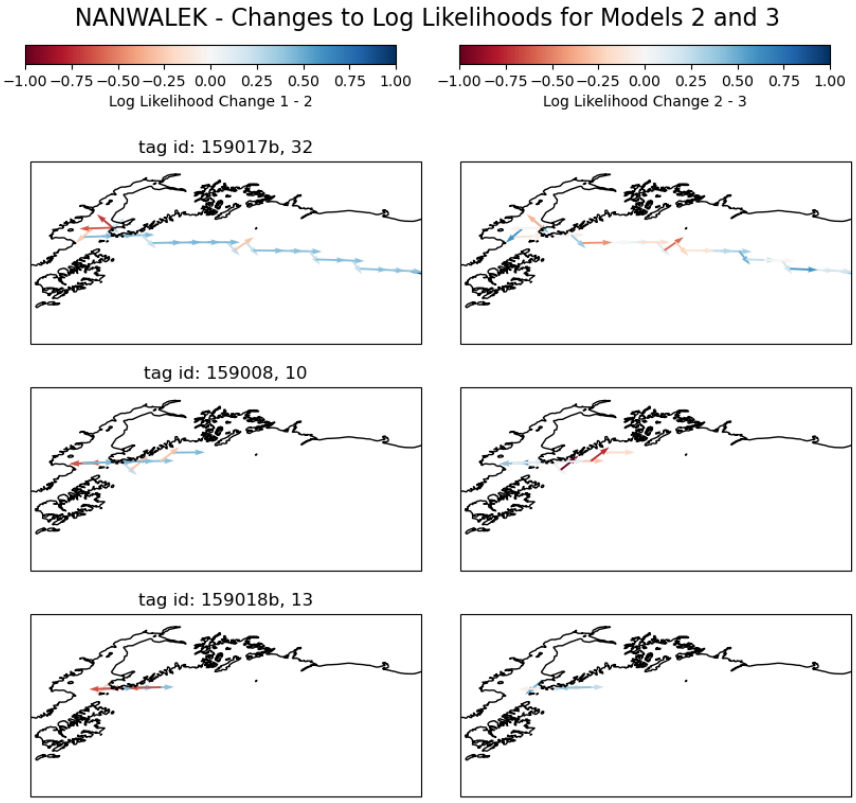

Figure 7: Log likehood changes per decision - Nanwalek

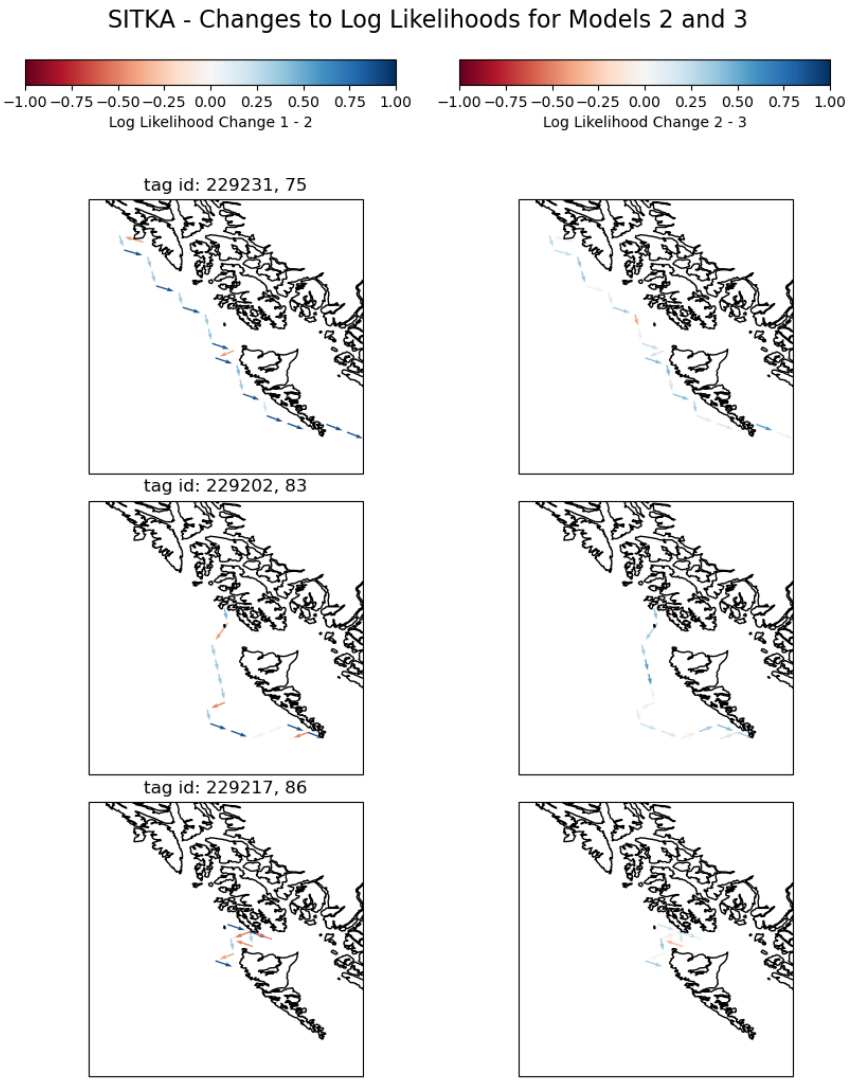

Figure 8: Log likehood changes per decision - Sitka

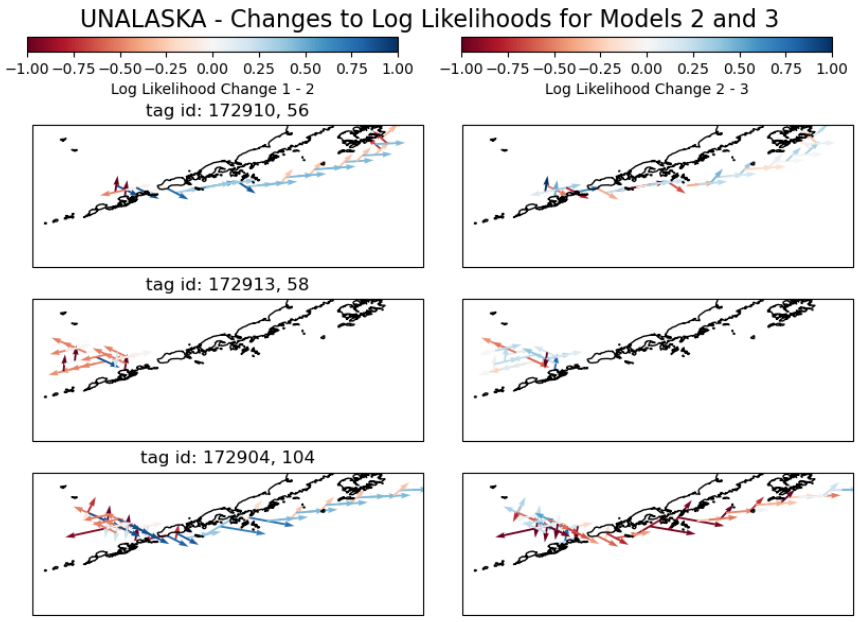

Figure 9: Log likehood changes per decision - Unalaska

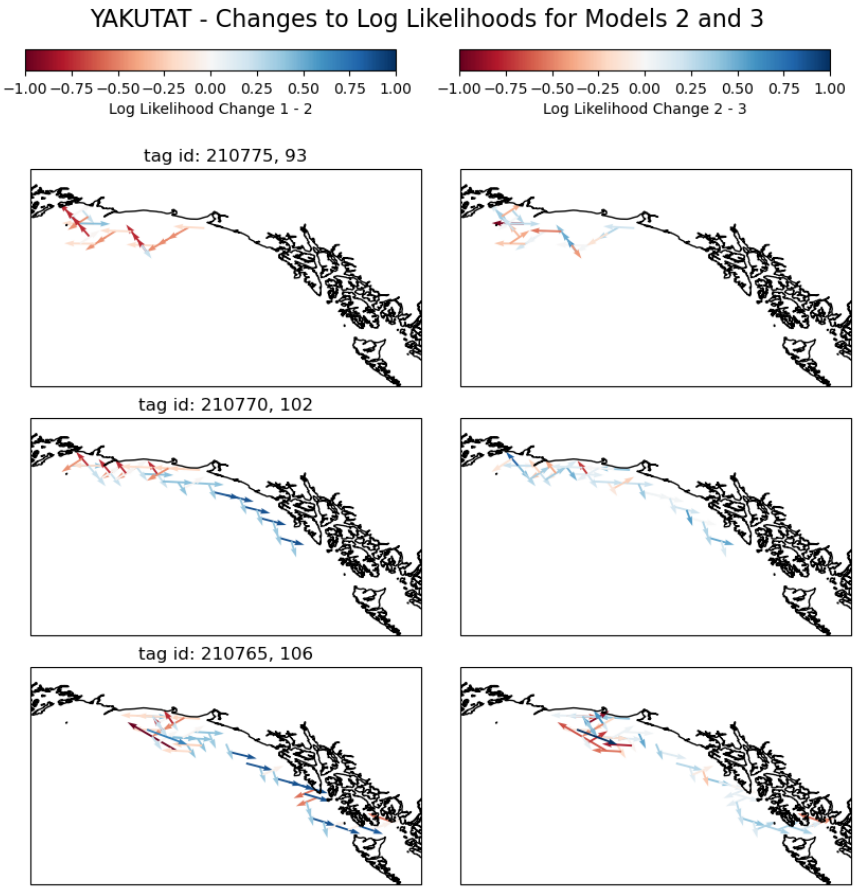

Figure 10: Log likehood changes per decision - Yakutat