Model-based approaches for analyzing fishery dependent CPUE

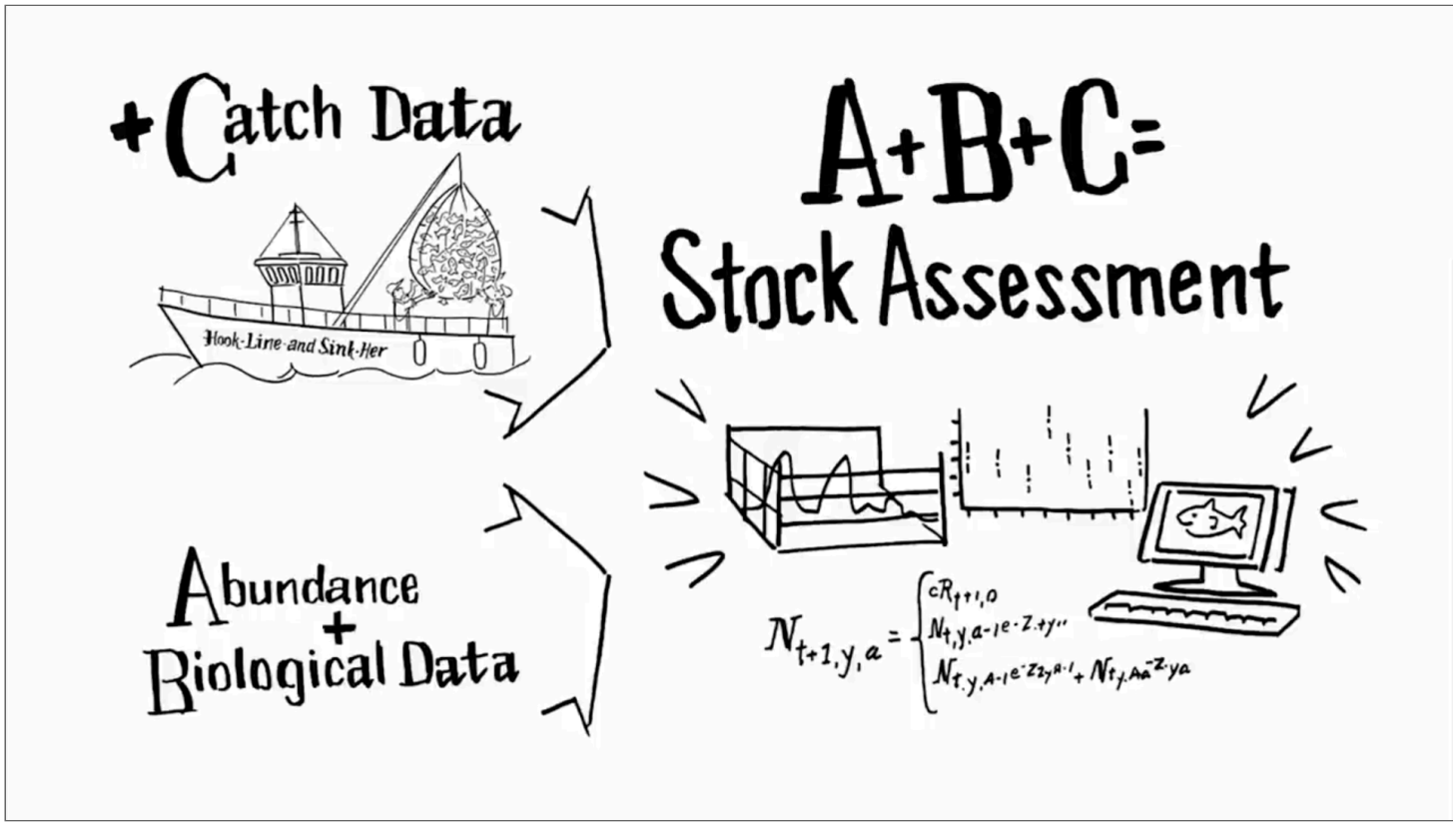
Model overview, simulation results, and extensions for index generation

Eric Ward, Kiva Oken, Allen Chen, Kate Richerson, Dan Holland, Cheryl Barnes

October 5, 2026

These materials do not constitute a formal publication and are for information only. They are in a pre-review, pre-decisional state and should not be formally cited or reproduced. They are to be considered provisional and do not represent any determination or policy of NOAA or the Department of Commerce.

Indices of abundance are critical building blocks of stock assessments



From “The ABCs of Stock Assessments” by NOAA Fisheries

Strength of information from West Coast Groundfish Bottom Trawl Survey varies by species

- Valuable source of biological data from full assessment area
- Extremely informative index for a subset of species
- Not generally used for nearshore stocks
- Index can be weakly informative for many rockfish species



How could we improve the strength of information of stock assessment indices from existing data?

1. Identify and control for time-varying drivers of availability of species to the bottom trawl survey
   - E.g., what drives the percent of time in sampled vs unsampled habitats?
2. Develop fishery-dependent indices from different data sources
   - eDNA
   - Acoustic trawl survey index for non-whiting groundfish
   - Combining hook-and-line surveys conducted by individual states
3. Utilize fishery-dependent data, control for non-random nature of the sampling
   - Recreational indices common in nearshore stock assessments
   - **Commercial fishery-dependent indices less common**

Plan for the rest of the day

Part I: theory, simulation

1. Theory of index standardization
2. Challenges and theoretical solutions for non-random (preferentially sampled) data
3. Simulation study design for preferentially sampled survey data
4. Simulation results, recommendations

Part II: application

More later

Why model-based index standardization?

- Catch at a tow reflects **density** × **catchability** × **where and when we sampled**
- Sampling is rarely balanced: strata may be missed, survey timing shifts, vessels change
- Spatiotemporal models let us
- borrow strength across space and time
- separate persistent hotspots from year-specific shifts
- predict density on a *fixed* grid every year
- NWFSC has a long history of advancing these methods

Model - based workflows at NWFSC

- Leverages R package `sdmTMB`

Delta (hurdle) observation model

Two components, for location s and year t (component 1: encounter; component 2: positive catch):

$$y_{s,t} \sim \text{Bernoulli}(p_{s,t}) \qquad \text{encounter}$$

$$y_{s,t} \mid y_{s,t} > 0 \sim \text{Gamma}(\text{shape} = \phi,\ \text{mean} = r_{s,t}) \qquad \text{positive catch}$$

$$\mathbb{E}[y_{s,t}] = p_{s,t}\, r_{s,t}$$

sdmTMB family	Positive-catch model
`delta_gamma()`	$y_{s,t} \mid y_{s,t} > 0 \sim \text{Gamma}(\phi,\ r_{s,t})$
`delta_lognormal()`	$\log y_{s,t} \mid y_{s,t} > 0 \sim \text{Normal}\left(\log r_{s,t} - \frac{\phi^2}{2},\ \phi^2\right)$

In both families $r_{s,t}$ is the mean of the positive catch.

Linear predictors

$$\text{logit}(p_{s,t}) = \beta_{1,t} + \omega_{1,s} + \epsilon_{1,s,t}$$

$$\log(r_{s,t}) = \beta_{2,t} + \omega_{2,s} + \epsilon_{2,s,t}$$

- **Year fixed effects** $\beta_{k,t}$: a separate mean for each year (`0 + as.factor(year)`);
- **Spatial field** $\omega_{k,s}$: persistent spatial pattern, shared by all years (habitat, etc)
- **Spatiotemporal field** $\epsilon_{k,s,t}$: year-specific deviations from that pattern (shifts, patchiness)

Subscript $k \in \{1, 2\}$ is the delta component; each gets its own parameters.

Gaussian random fields

$$\boldsymbol{\omega}_k \sim \text{MVN}(\mathbf{0},\ \boldsymbol{\Sigma}_{\omega,k}), \qquad \boldsymbol{\epsilon}_{k,t} \sim \text{MVN}(\mathbf{0},\ \boldsymbol{\Sigma}_{\epsilon,k})$$

Matérn covariance ($\nu = 1$)

$$\Sigma_{ij} = \sigma^2\, \kappa d_{ij}\, K_1(\kappa d_{ij})$$

- marginal SDs: σ_O (spatial), σ_E (spatiotemporal)
- range $= \sqrt{8}/\kappa$: distance where correlation ≈ 0.13
- SPDE on a triangulated mesh -> sparse precision matrix $\boldsymbol{Q}$

Structure of $\boldsymbol{\epsilon}_{k,t}$ over time

- `"iid"`: independent each year
- `"ar1"`: $\boldsymbol{\epsilon}_t = \rho\, \boldsymbol{\epsilon}_{t-1} + \boldsymbol{\delta}_t$
- `"rw"`: $\rho = 1$
- `"off"`: spatial only

From model to index

$$\hat{I}_t = \sum_{j=1}^{n_j} A_j \, \hat{p}_{j,t} \, \hat{r}_{j,t}$$

- j: prediction grid cells covering the survey domain;
- A_j: cell area
- $\text{logit}(p_{s,t})$ probability of occurrence
- $\log(r_{s,t})$ log expected catch

Density vs. catchability

Add terms that change **catch** but not **density**, then predict with them held at a reference value. For tow i at location $s[i]$ in year $t[i]$:

$$\log(r_i) = \underbrace{\beta_{2,t[i]} + \omega_{2,s[i]} + \epsilon_{2,s[i],t[i]}}_{\text{density}} + \underbrace{f_2(d_i) + \alpha_{2,v[i]} + \boldsymbol{x}_i^\top \boldsymbol{\gamma}_2}_{\text{catchability}}$$

- $f(d_i)$: seasonality / day of year
- $\alpha_{v[i]}$: vessel (or vessel-year) random effects
- $\boldsymbol{x}_i^\top \boldsymbol{\gamma}$: other factors (gear, time of day, ...)

Same structure applies to $\text{logit}(p_i)$.

Preferential sampling

What is preferential sampling?

- Non-random distribution of sampling locations in space or time
- Locations can depend on density or other factors
- Common in fishery-dependent data
- If ignored, the model sees mostly high-density places, which can bias the index

Joint (density-based)
Selection depends on density (Diggle et al. 2010; Conn et al. 2017)

Anomaly-based (discrete choice)
Selection depends on a local catch anomaly that also inflates catch

Next: can spatial models absorb this, and how can we correct for it?

Can spatial models absorb preferential sampling?

- Ducharme-Barth et al. (2022) *Fisheries Research* 246:106169
- Preferential sampling: $P_x \propto S_x^{\phi}$, baseline $\phi = 0.5$ (sampling $\propto \sqrt{\text{abundance}}$)
- Sensitivity: $\phi \in \{0, \ldots, 8\}$; closures; fleet expansion and contraction

Why spatial models looked sufficient

- ϕ was constant over time: no change in preference strength
- Time variation came from the footprint (closures, contraction), not preference
- High sample sizes (60K), low observation error (CV = 0.15)

Their recommendation: non-stationarity could still cause bias and suggested joint models, which is part of our focus.

Correcting for preferential sampling

We explore two ways to correct the bias

	Joint sampling model	Discrete choice model
Origin	Diggle et al. (2010); Conn et al. (2017)	Roy (1951); Dahl (2002); Chen et al. (2023)
Unit of the sampling model	every grid cell: sampled or not, $R_{j,t}$	each tow: chosen cell vs. nearby candidates
Sampling model	Bernoulli, $\text{logit}(\pi_{j,t}) = \gamma_0 + b_t\,\delta_{j,t} + \eta_j$	conditional logit, $V_{s,t} = b_t\,z_{s,t} + \eta_s$
Link to catch	shared density field $\delta_{j,t}$	correction term added to positive catch
Assumes effort follows	expected density**	expected utility, incl. private information
Preferential strength	b_t, random walk	b_t, random walk

In both models the index uses only the density terms: η and the correction terms are left out.

The Conn et al. (2017) joint model

Model sampling locations alongside what was caught, linked by the same density field.

$$\text{Catch:} \quad y_i \sim \text{delta model with } \log r_{s,t} = \beta_{2,t} + \omega_{2,s} + \epsilon_{2,s,t}$$

$$\delta_{s,t}$$

$$\text{Sampling:} \quad R_{j,t} \sim \text{Bernoulli}(\pi_{j,t}), \qquad \text{logit}(\pi_{j,t}) = \gamma_0 + b\,\delta_{j,t} + \eta_j$$

- $R_{j,t} = 1$ if grid cell j was sampled in year t, over every cell in the domain
- The shared $\delta_{j,t}$ is the link: $b > 0$ means sampled cells tend to be high-density ones
- Cells that were never sampled are informative too
- $b = 0$ recovers the naive model; the index still uses only the density terms

Builds on Diggle et al. (2010) and Pati et al. (2011). Here we extend b to a random walk b_t.

The latent sampling field η

$$\boldsymbol{\eta} \sim \mathrm{MVN}(\mathbf{0},\ \boldsymbol{\Sigma}_\eta)$$

- Its **own** Matérn parameters: marginal SD σ_η and range ρ_η
- **Independent of density**: no correlation with ω or ϵ
- Constant across years

What it represents

- Distance to port, fuel cost, travel time
- Closed areas and regulations
- Untrawlable bottom etc
- Fleet habits, personal preferences

Without η, any clustering of effort would be interpreted as density-driven, biasing b. With η, b captures only the density-related part of where we sample.

Density and sampling fields together

- Sampling concentrates where density is high and where the sampling field is high
- η is not in the index: it doesn't change how many fish are there
- In the simulation: η_j drawn once with `sdmTMB_simulate()` (range 25 km, $\sigma = 1$), then standardized

Discrete choice: modeling location choice

Each tow i is a choice of cell s_i among a candidate set $\mathcal{C}_i$ in year t_i:

$$V_{s,t} = b_t \, z_{s,t} + \eta_s, \qquad p_i = \frac{\exp\{V_{s_i,t_i}\}}{\sum_{s' \in \mathcal{C}_i} \exp\{V_{s',t_i}\}}$$

- $z_{s,t}$: expected log density from the density model
- b_t: preferential strength, random walk over years
- η_s: latent selection field (in economics models: distance, revenue, cost)
- $\mathcal{C}_i$: the chosen cell plus candidate cells
- Choice likelihood: $\prod_i p_i$, multiplied into the joint likelihood with the catch model

Discrete choice: correcting the catch model

Fishers choose a cell when its unobserved catch deviation is positive, so $E[\varepsilon_i \mid \text{chose } s_i] \neq 0$.

$$\log \hat{\mu}_{2,i}^{\text{corr}} = \log \hat{\mu}_{2,i} + \eta_{s_i} + \lambda_{0,s_i} + \lambda_{1,s_i}\, p_i + \lambda_{2,s_i}\, p_i^2$$

- High p_i: chosen for observable reasons, little selection bias
- Low p_i: chosen *despite* low expected catch, positive selection bias
- The polynomial in p_i approximates this (control function; Chen et al. 2023)
- $\lambda_0, \lambda_1, \lambda_2$: spatially varying coefficients, each a GRF
- Separate SDs, one shared (fixed) range
- Applied to the positive component only; encounter uncorrected
- Previous applications estimated λ separately per cell

Simulating preferentially sampled data

Simulation workflow

For each year t and grid cell j:

1. **Truth**: density $D_{j,t} = p_{j,t}\, r_{j,t}$ from a delta model fit to real data
2. **Sampling field**: a non-density spatial field η_j (e.g. access, fishing habits, port proximity)
3. **Selection**: probability $\pi_{j,t}$ depends on η_j and on density through b
4. **Draw locations**: n tows per year with probability $\propto \pi_{j,t}$
5. **Simulate catch** at the chosen cells from the delta model

- We use 10 years of data from the WCGBTS for shortspine thornyhead (2012-2022) to anchor the simulations to real data (and sdmTMB models used for index standardization)

Steps 1–2: truth and a sampling field

Density covariate

$$\tilde{D}_{j,t} = \frac{\log D_{j,t} - \overline{\log D}}{\mathrm{sd}(\log D)}$$

standardized over all cells *and* years

Sampling field (not density)

$$\eta_j \sim \mathrm{GRF}(\mathrm{range} = 25\ \mathrm{km},\ \sigma = 1)$$

standardized; fixed across years and replicates

Steps 3–4: selecting where to sample

$$\text{logit}(\pi_{j,t}) = \gamma_0 + \gamma_\eta \, \eta_j + b \, \tilde{D}_{j,t}$$

- b constant over time: $b_t = b$ for all t
- $b = 0$: random-sampling control
- $b > 0$: log density multiplies the odds of sampling
- $n = 1000$ cells per year, drawn with replacement, weights $\propto \pi_{j,t}$

Symbol	Code	Value
γ_0	`c0_samp`	-2
γ_η	`gamma_samp_true`	1
b	`b_max` / `b_pref`	0, 3, 5

Where the samples land

- Toy example: same density surface and sampling field, 200 draws per panel
- Only relative weights matter, raising b from 3 to 5 changes little

Step 5: simulating catch

At each selected cell, observations come from the *true* density, not the selection model:

$$z_i \sim \text{Bernoulli}(p_{j[i],t})$$

$$y_i \mid z_i = 1 \sim \text{Lognormal}\left(\log r_{j[i],t} - \tfrac{\sigma^2}{2},\ \sigma\right), \qquad \sigma = \sqrt{\log\left(1 + \text{CV}^2\right)}$$

- $\text{CV} = 1/\sqrt{\phi}$ from the fitted Gamma, so the lognormal matches its variability
- The sampling indicator $R_{j,t} = 1$ if cell j was sampled in year t is kept for the joint model

> Constant b biases every year about equally, so we expect mostly a **level** bias in the naive index, not a spurious trend.

Time-varying preferential sampling

Why let b change over time?

- Fleets learn: better technology, shared information
- Economics and management shifts (prices, closures, quotas)
- The share of tows in hotspots changes from year to year

Constant b
Sampling is equally biased every year, so we expect mostly a level bias

Increasing b_t
Effort concentration changes relative to density, so the naive index can show a spurious trend

Replace b with b_t in both data-generating approaches.

Two ways to simulate preferential data

Density-based (joint)

$$\text{logit}(\pi_{j,t}) = \gamma_0 + \gamma_\eta \eta_j + b_t \, \tilde{D}_{j,t}$$

- Selection depends on true density
- Catch simulated from the true density
- Estimated with a joint (Conn-style) sampling model

Anomaly-based (discrete choice)

$$\text{logit}(\pi_{j,t}) = \gamma_0 + \gamma_\eta \eta_j + b_t \, A_j$$

$$\log r^*_{j,t} = \log r_{j,t} + \theta A_j$$

- A_j: standardized catch anomaly field (GRF, range 40 km, $\sigma = 1$), $\theta = 1$
- The anomaly drives both selection and positive catch
- Estimated with a discrete-choice model

In both cases tows are drawn with probability $\propto \pi_{j,t}$, $n = 1000$ per year; η_j is the same sampling field as before.

The b_t process: linear trend + AR(1) deviations

$$\mu_t = b_{\max} \frac{t-1}{T-1}$$

$$d_t = \phi\, d_{t-1} + e_t, \quad e_t \sim \text{Normal}(0, \sigma_b^2), \quad d_1 \sim \text{Normal}\left(0, \frac{\sigma_b^2}{1-\phi^2}\right)$$

$$b_t = \max(0,\ \mu_t + d_t)$$

- μ_t: linear ramp from 0 in the first year to $b_{\max}$ in the last
- d_t: stationary AR(1) deviations, $\phi = 0.7$, $\sigma_b = 0.3$
- Lower bound at 0
- Unique trajectory for each replicate (2012–2022)

Example b_t trajectories

Illustrative draws from the same process.

Simulation design

Component	Setting
Truth	Delta-gamma sdmTMB fit to WCGBTS shortspine thornyhead, 2012–2022, 4 km² grid
Preferential strength	$b_{\max} \in \{0, 3, 5\}$; $b_{\max} = 0$ is the random-sampling control
b regime	constant ($b_t = b_{\max}$) or time-varying (trend + AR(1))
Data-generating model	density-based or anomaly-based
Replicates	25 datasets per scenario, 1000 tows per year, 11 years
Seeds	`set.seed(rep * 100)`: replicate r uses the same seed in every scenario
Grids for the index	full survey grid, and a 10 km buffer around each year's tows

Every scenario is compared against the truth and against the $b_{\max} = 0$ control fit to the same replicate.

Simulation workflow

1. Truth

Fit sdmTMB to WCGBTS data; predict $p_{j,t}$, $r_{j,t}$ on the grid

→

2. Fixed fields

Draw η_j and A_j once; same in every replicate

→

3. Simulate

For each $b_{\max}$ and replicate: draw b_t, select tows, simulate catch

→

4. Estimate

Fit naive, joint and discrete-choice models

→

5. Compare

Index bias and trend (full and buffer grid); recovered b_t

Estimation model	Sampling model	Preferential term	Fit to
Naive	none	none	all scenarios
Joint	Bernoulli $R_{j,t}$ on every grid cell	random walk b_t	density-based, $b_{\max} > 0$
Discrete choice	chosen cell vs. candidates	random walk b_t	anomaly-based, $b_{\max} > 0$

Both preferential estimation models use a random walk, not the trend + AR(1) used to generate the data.

Results

Constant b: a level bias

- Density-driven data, constant b = 0, 3, 5; naive model
- Stronger preferential sampling shifts the index up
- The trend is essentially unchanged
- A level offset is absorbed by catchability (q) in most assessments

Mean and 95% CI across replicates.

Time-varying b_t: estimated indices

Indices scaled to the first year. Rows: data-generating mechanism; columns: $b_{\max}$. Dashed line is the truth.

Error in level and trend

- **Naive**: trend overstated by about 9% (density-driven) and 30% (anomaly-driven)
- **Joint**: lower RMSE and trend error under both mechanisms
- **Discrete choice**: best or near-best when data generation is anomaly-driven, can be worse than naive when density-driven

Top: index RMSE; bottom: slope error (estimate − truth). Mean and 95% CI across 25 replicates.

How much trend bias is removed?

Mechanism	$b_{\max}$	Naive slope bias	Joint	Discrete choice
Density-driven	3	+0.0036	76% removed	bias ×3 (+0.0116)
Density-driven	5	+0.0040	89% removed	bias ×2 (+0.0087)
Anomaly-driven	3	+0.0137	66% removed	75% removed
Anomaly-driven	5	+0.0117	99% removed	64% removed

- True log-linear slope ≈ 0.042 per year; $b_{\max} = 0$ naive bias ≈ −0.002
- Joint model converged in 19–23 of 25 replicates; naive and discrete choice in all 25

The joint model helps regardless of mechanism; discrete choice only helps when the mechanism matches it.

Why discrete choice over-corrects

- Correction comes almost entirely from the latent field η_s, not the choice-probability polynomial
- Polynomial coefficients ($\lambda_0, \lambda_1, \lambda_2$) stayed near their starting values across replicates
- **Anomaly-driven**: η_s recovers the unobserved catch anomaly and fixes the index
- **Density-driven**: η_s soaks up real spatial pattern in density, pushing the trend further from the truth

$$V_{s,t} = \alpha_t \, z_{s,t} + \eta_s \qquad \log \hat{\mu}_{2,i}^{\text{corr}} = \log \hat{\mu}_{2,i} + \eta_{s_i} + \lambda_{0,s_i} + \lambda_{1,s_i} p_i + \lambda_{2,s_i} l$$

Consequences through the lens of the assessment

Shortspine assessment includes:

- Catches
- Disccard rates
- Lengths
- Indices
- Growth, natural mortality are fixed

Consequences through the lens of the assessment

Assessment model fit to original survey index

Recruitment time series

High recruitment in 2000s drives assessment model fit of the monotonically increasing survey index

Consequences through the lens of the assessment

Simulated indices added to the 2023 shortspine thornyhead assessment as a new fleet

$b_{\max} = 5$	Naive	Joint	Baseline
R_0	12,624	12,523	12,551
2008 rec dev (ε_{2008})	1.17	1.08	1.05
OFL_{2025}	927	911	919
OFL_{2034}	1,160	1,140	1,139

Medians across replicates. Baseline: naive index with $b = 0$ (red line).

Take-aways

Take-aways

1. Constant preferential sampling mainly biases the level of an index, which catchability absorbs
   a. off the shelf tools, like `sdmTMB` / `VAST` are fine in these situations
2. Time-varying preferential sampling biases the trend, which catchability cannot fix
   a. off the shelf tools, like `sdmTMB` / `VAST` will not save you
3. The joint model reduced trend bias under both density- and anomaly-driven sampling
4. The discrete-choice model worked well when sampling followed unobserved catch anomalies, but made the trend worse when sampling tracked density
5. Uncorrected trend bias carries into the assessment: an inflated major recruitment event and overfishing limits whose bias grows in projections

In practice: Specifying the sampling model

- Standardize the density term in the sampling model
- Why: reduces correlation between b and γ_0
- Use the spatial deviations (ω + ϵ) and ignore year effects
- Why: otherwise changes in abundance year-to-year affects how much sampling happens
- For real world datasets, consider giving intercept γ_0 same flexibility as slope
- For real world datasets, consider giving η its own spatial range
- Allows fish patchiness and fishing patchiness to work at different scales

In practice: Fitting and diagnostics

- Check convergence carefully
- Maximum gradient, Hessian, a la sdmTMB `sanity`
- Examine CIs on estimated b_t trajectory
- Always compare model with time-varying preferential sampling to naive model
- When mechanism is unknown, treat joint model as the default

In practice: limitations and extensions

- Spatial pattern of the sampling field η_s fixed over time
- Fixed patchiness (range) of fishing effort
- No schooling or shoaling behavior

What's next & future directions

- Evaluating joint sampling and discrete choice models applied to data from the midwater trawl sector
- extensions including fishing and environmental covariates
- more realistic discrete choice models
- continuing to evaluate impacts on stock assessments (for 2029 and beyond)
- Methods described here are applicable to many other sectors - species in the USA
- teaser: currently examining midwater hake indices, etc

Speaker notes