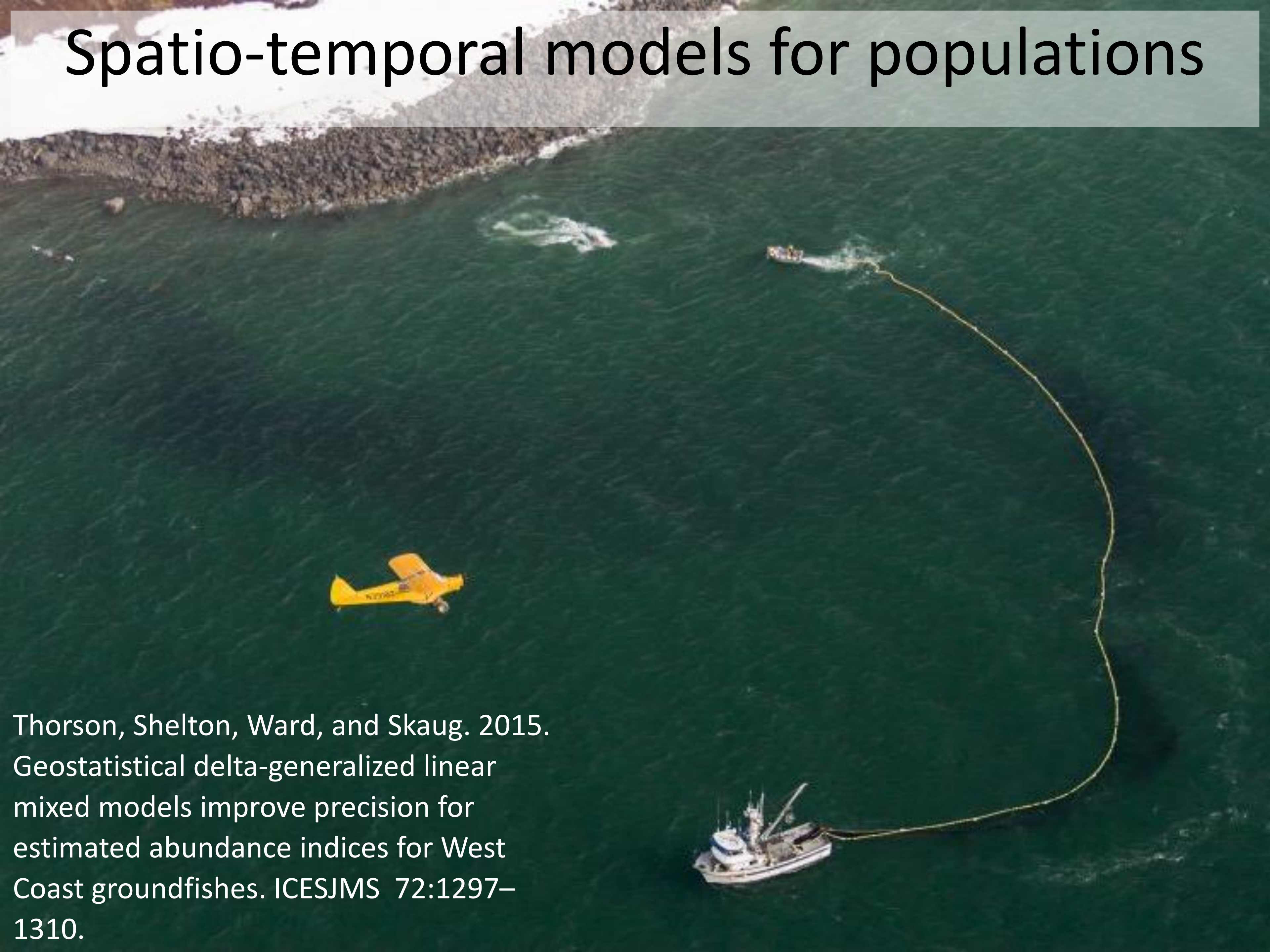
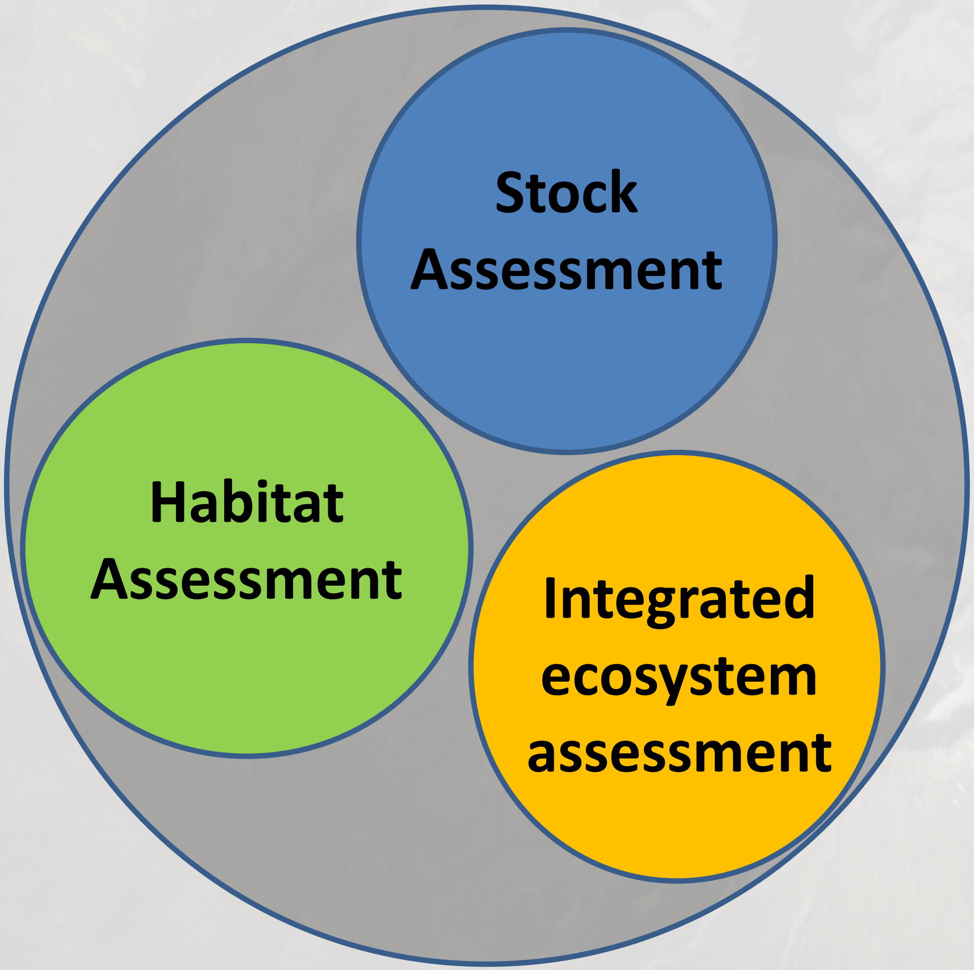
# Spatio-temporal models for populations



Thorson, Shelton, Ward, and Skaug. 2015. Geostatistical delta-generalized linear mixed models improve precision for estimated abundance indices for West Coast groundfishes. ICESJMS 72:1297–1310.

## Spatio-temporal model

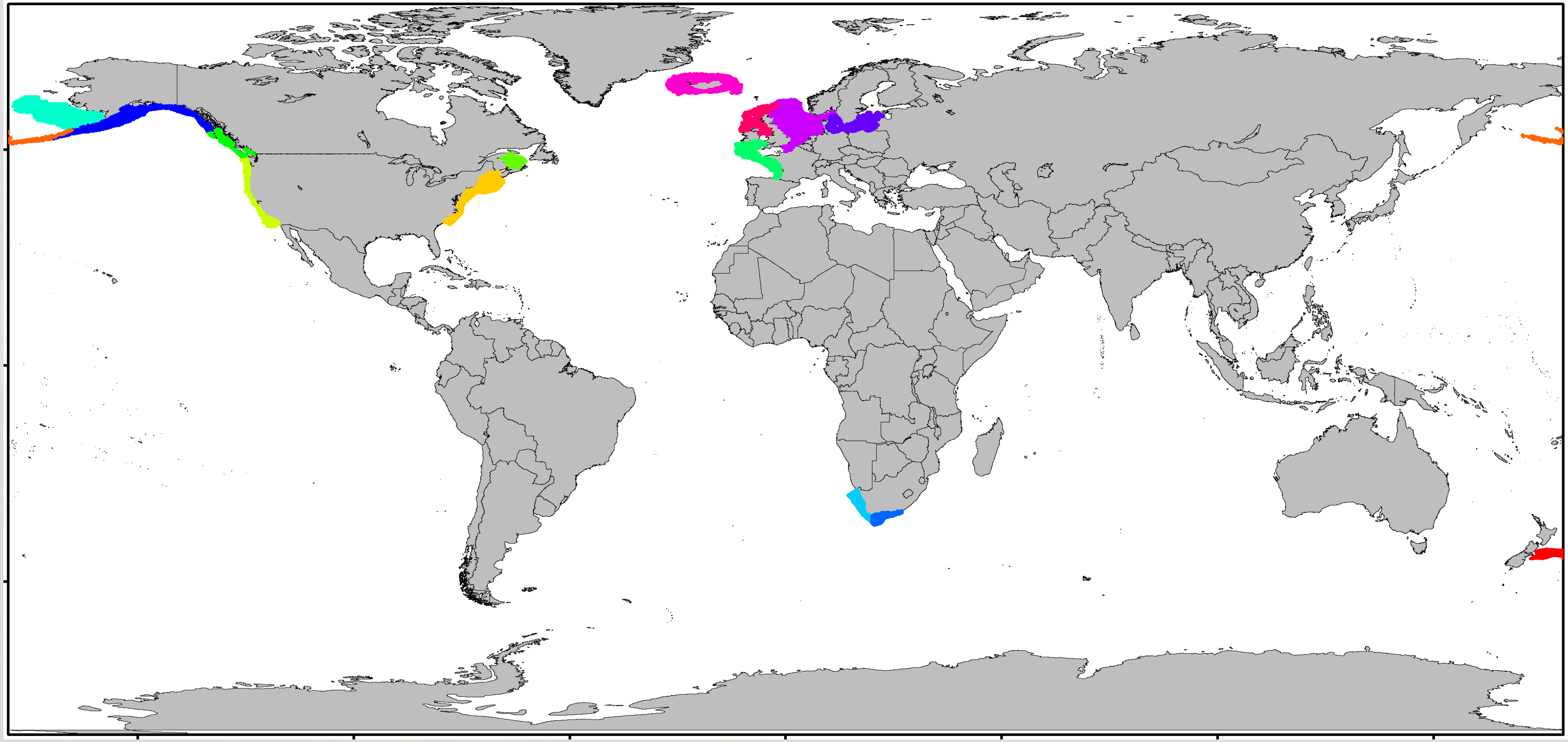



## Benefits of single approach

1. Include biological mechanism
2. Improved communication
3. Similar review standards and "burden of proof"

# Has been applied to >15 regions worldwide

```
> devtools::install_github("james-thorson/FishData")
Downloading GitHub repo james-thorson/FishData@master
from URL https://api.github.com/repos/james-thorson/FishData/zipball/
Installing FishData
```



# Four questions

- How should we impute density in areas with little data?
- When can we use auxiliary data to separate changes in fishery catchability and fish density?
- How should we account for non-random selection of fishing locations?
- How should we process “biological data” in conjunction with fishery CPUE?

# Four questions

- **How should we impute density in areas with little data?**
- When can we use auxiliary data to separate changes in fishery catchability and fish density?
- How should we account for non-random selection of fishing locations?
- How should we process “biological data” in conjunction with fishery CPUE?

# Delta-generalized linear mixed model (Delta-GLMM)

- Delta-model for observations

$$\Pr(B=b)=\begin{cases} 1-\gamma(s,t) & \text{if } B=0 \\ \gamma(s,t)\times g(B;\lambda(s,t)) & \text{if } B>0 \end{cases}$$

  - Where $\gamma(s,t)$ is the probability of encountering the species
  - $g(B;\lambda(s,t))$ is a distribution for positive catches
- Spatio-temporal variation in encounter probability

$$\text{logit}(\gamma(s,t))=\alpha_\gamma(t)+\omega_\gamma(s)+\varepsilon_\gamma(s,t)$$

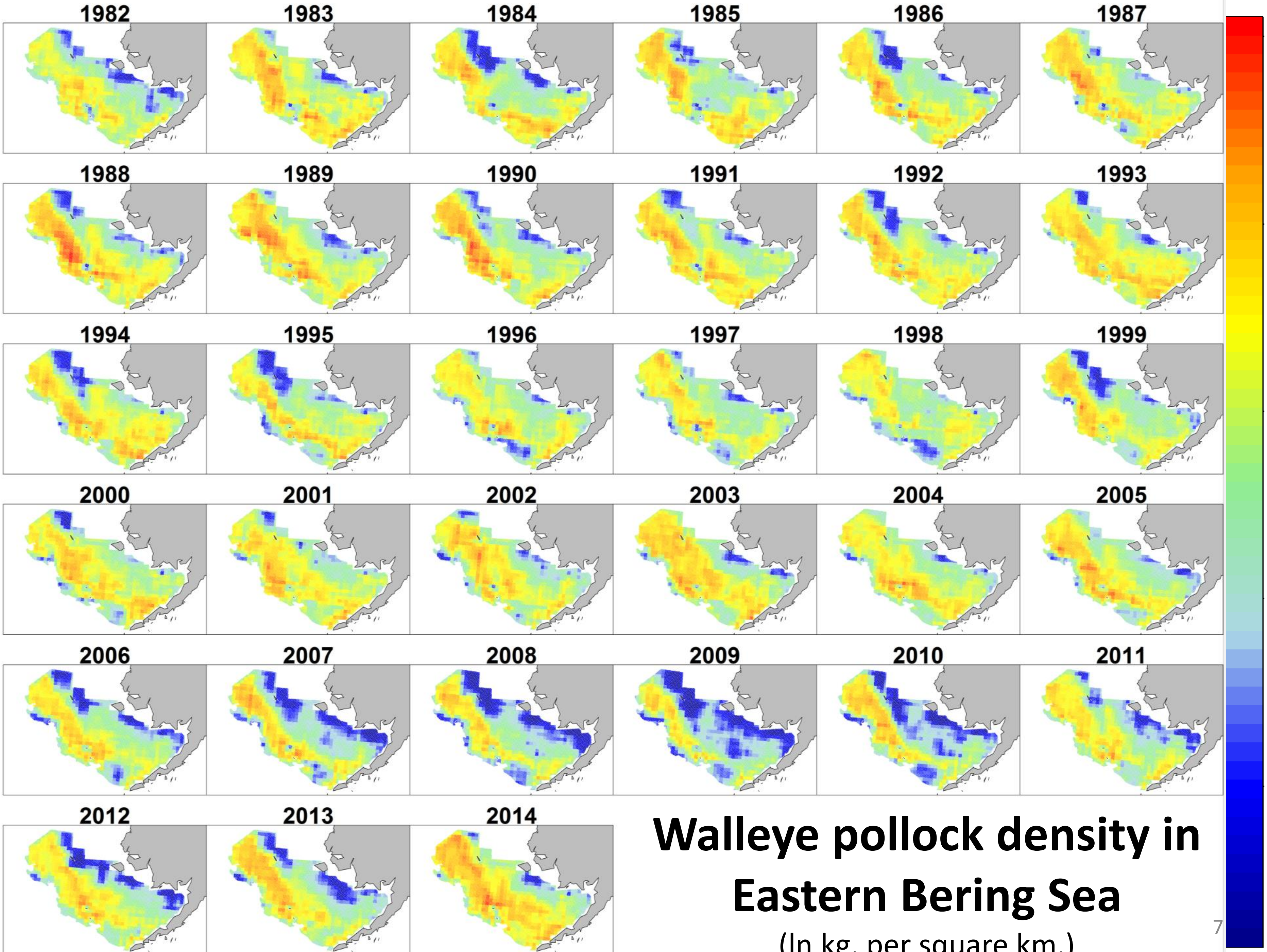
  - $\alpha_\gamma(t)$ is the intercept for each year
  - Where $\boldsymbol{\omega}_\gamma$ and $\boldsymbol{\varepsilon}_\gamma(t)$ follow a spatial distribution
- Spatio-temporal variation in density

$$\log(\lambda(s,t))=\alpha_\lambda(t)+\omega_\lambda(s)+\varepsilon_\lambda(s,t)$$

  - Where parameters are defined similarly to $\gamma(s,t)$
- Used to predict local density

$$\hat{d}(s,t)=\hat{\gamma}(s,t)\times\hat{\lambda}(s,t)$$

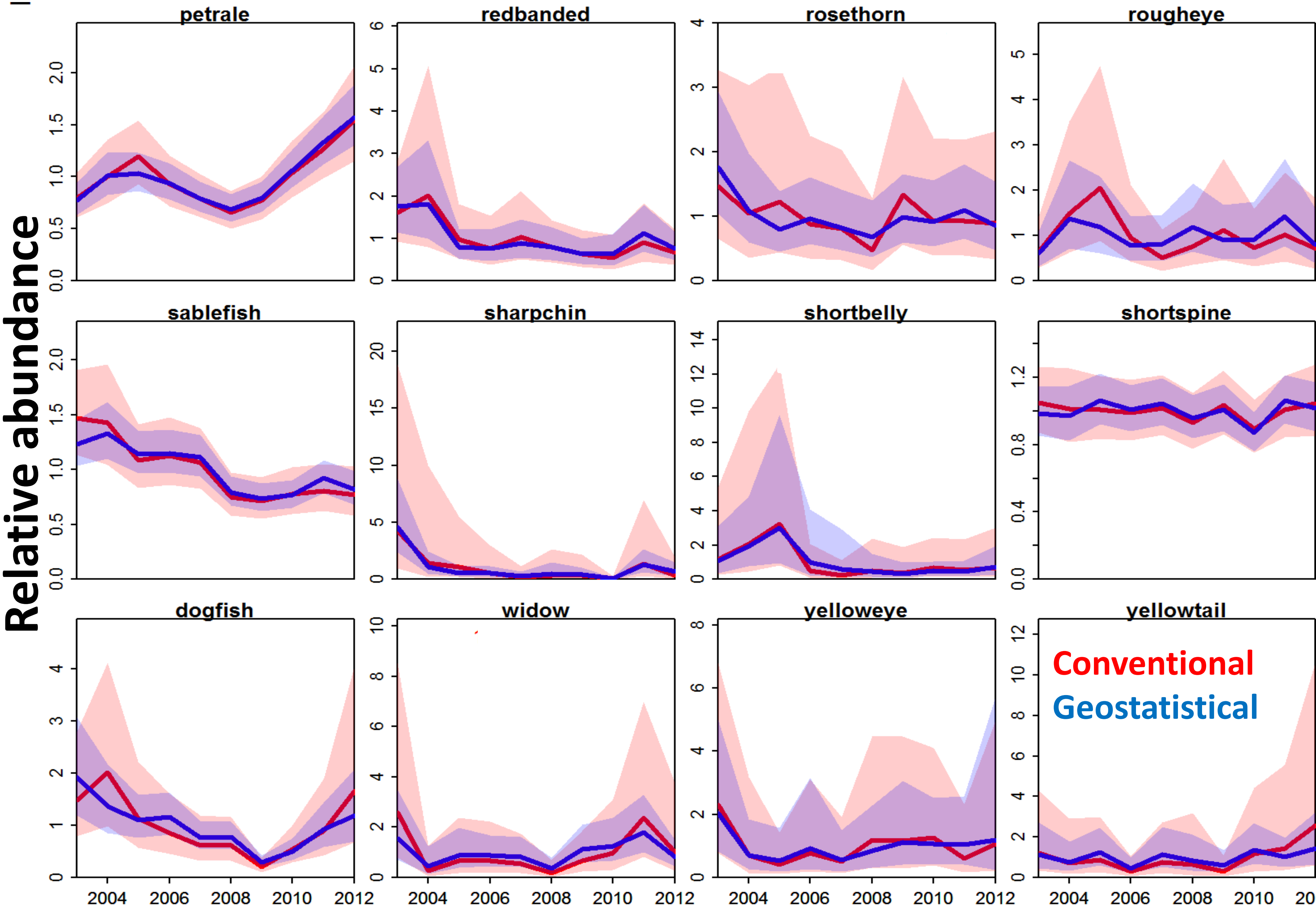
  - Where $\hat{\gamma}(s,t)$ and $\hat{\lambda}(s,t)$ are predictions conditioned on data




# Walleye pollock density in Eastern Bering Sea

(ln kg. per square km.)

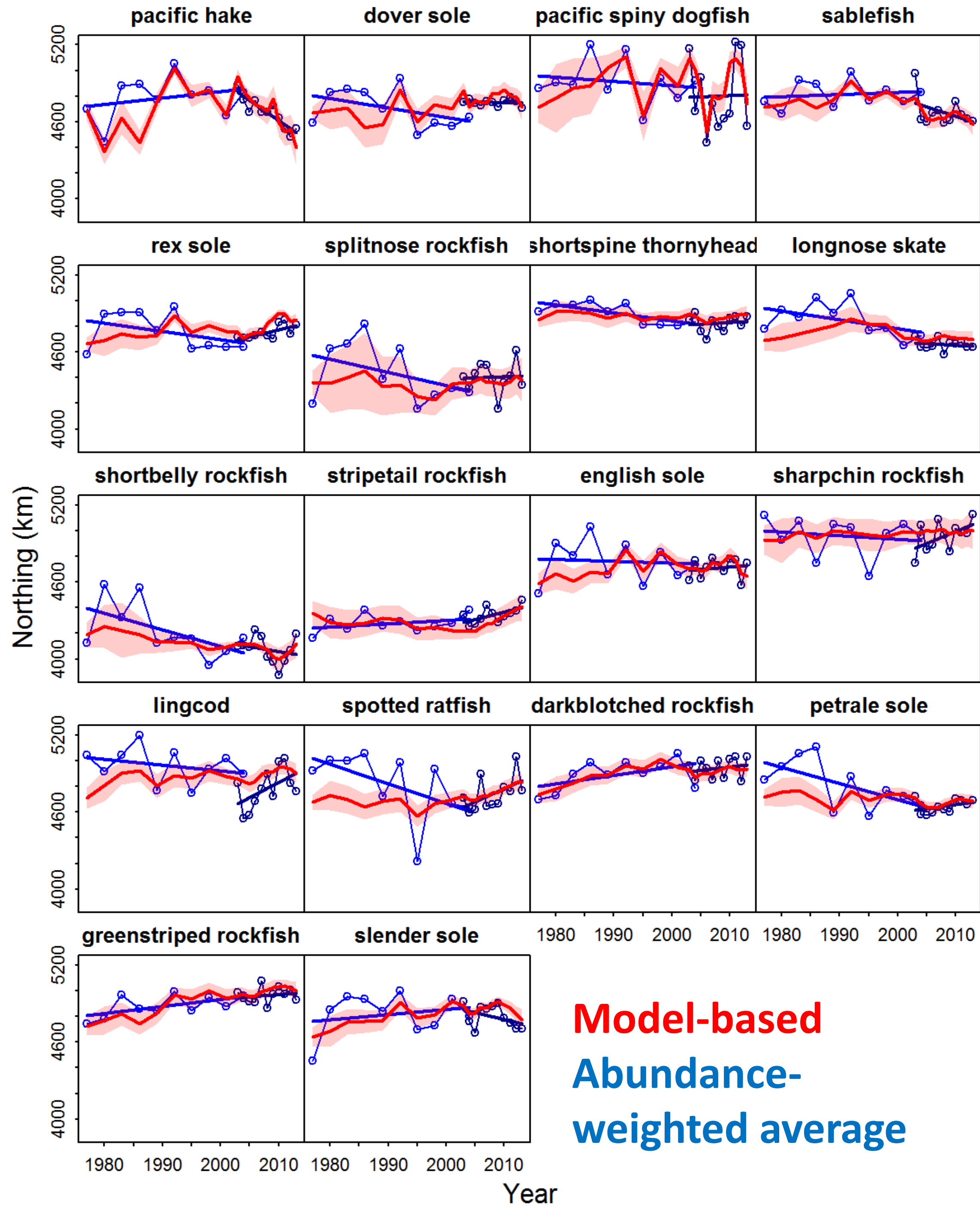
# Abundance indices



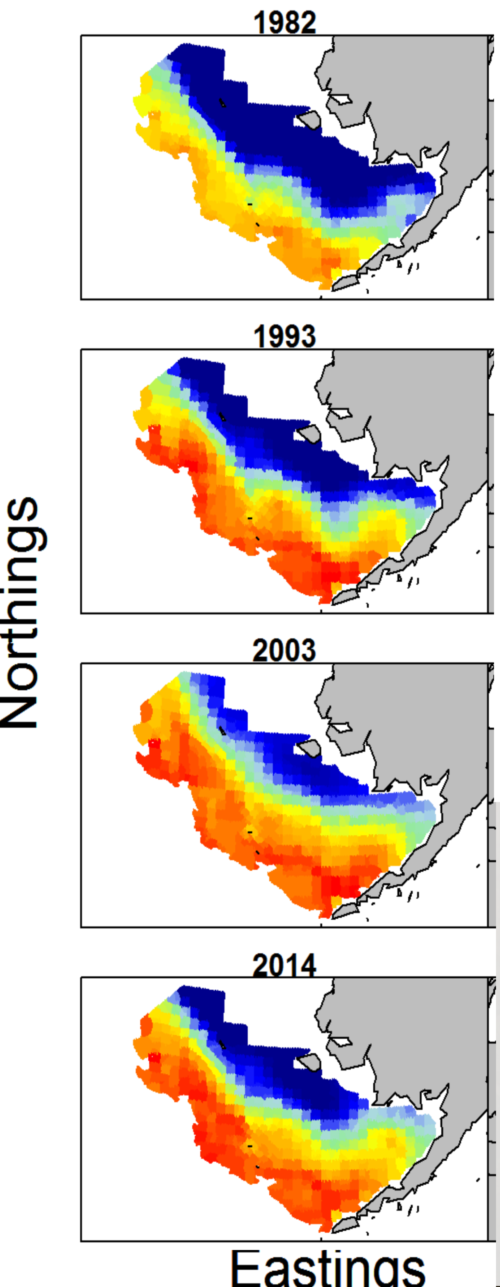
# Distribution shifts

- Highly variable distribution for semi-pelagic species
  - Dogfish
  - Sablefish
  - Hake
- Few clear trends
  - Depends on time-scale

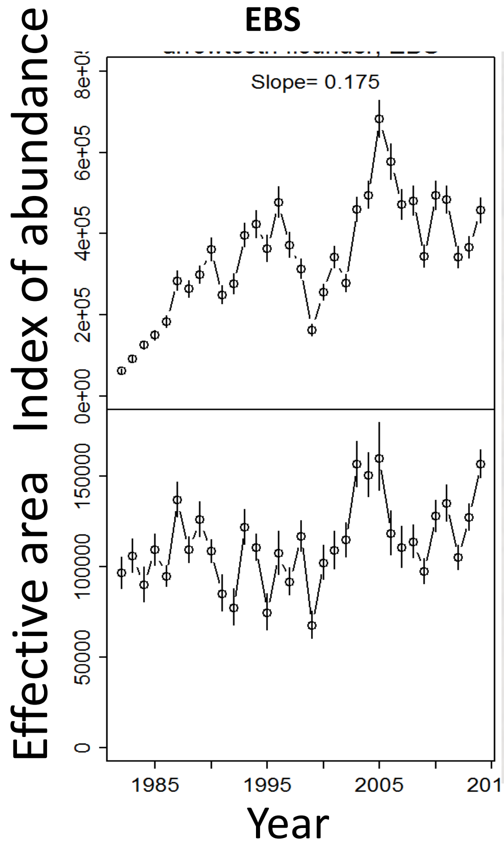
Thorson, Pinsky, and Ward. 2016. Model-based inference for estimating shifts in species distribution, area occupied and centre of gravity. Methods Ecol. Evol. **7**(8): 990–1002.



**Arrowtooth flounder**
**Eastern Bering Sea**




**Arrowtooth flounder**
**EBS**




Thorson, Rindorf, Gao, Hanselman, and Winker. 2016. Density-dependent changes in effective area occupied for sea-bottom-associated marine fishes. Proc R Soc B **283**(1840).

# Density-dependent habitat selection

- Do populations shrink their range when abundance is low?
- Average
  - Small contraction in range
  - Greatest in Eastern Bering Sea

# *Spatial Correlation*

**Sparse spatial correlation matrices**

- SPDE approximation
- 2D autoregressive process
- Stream network as Ornstein-Uhlenbeck process

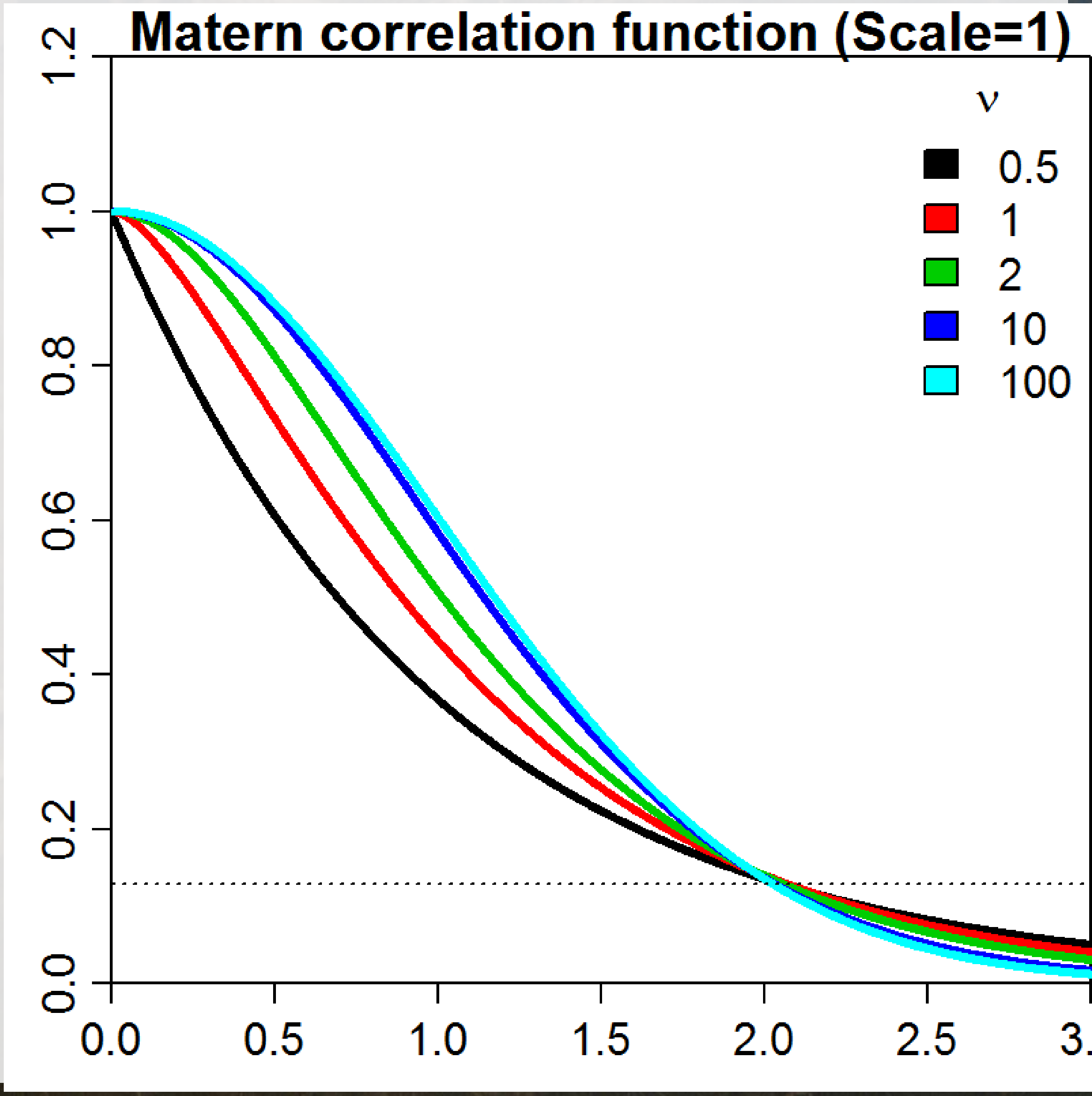
**Parameter estimation**

- Maximum marginal likelihood
  - Can use "bias-correction" for empirical Bayes predictions
- Template Model Builder
  - Automatic differentiation
  - Laplace approximation

# Spatial Correlation

Matern correlation function

- $\nu = 0.5$
  - Approximately exponential
- $\nu \to \infty$
  - Approximately Gaussian
- Differentiable $\lceil \nu - 1 \rceil$ times

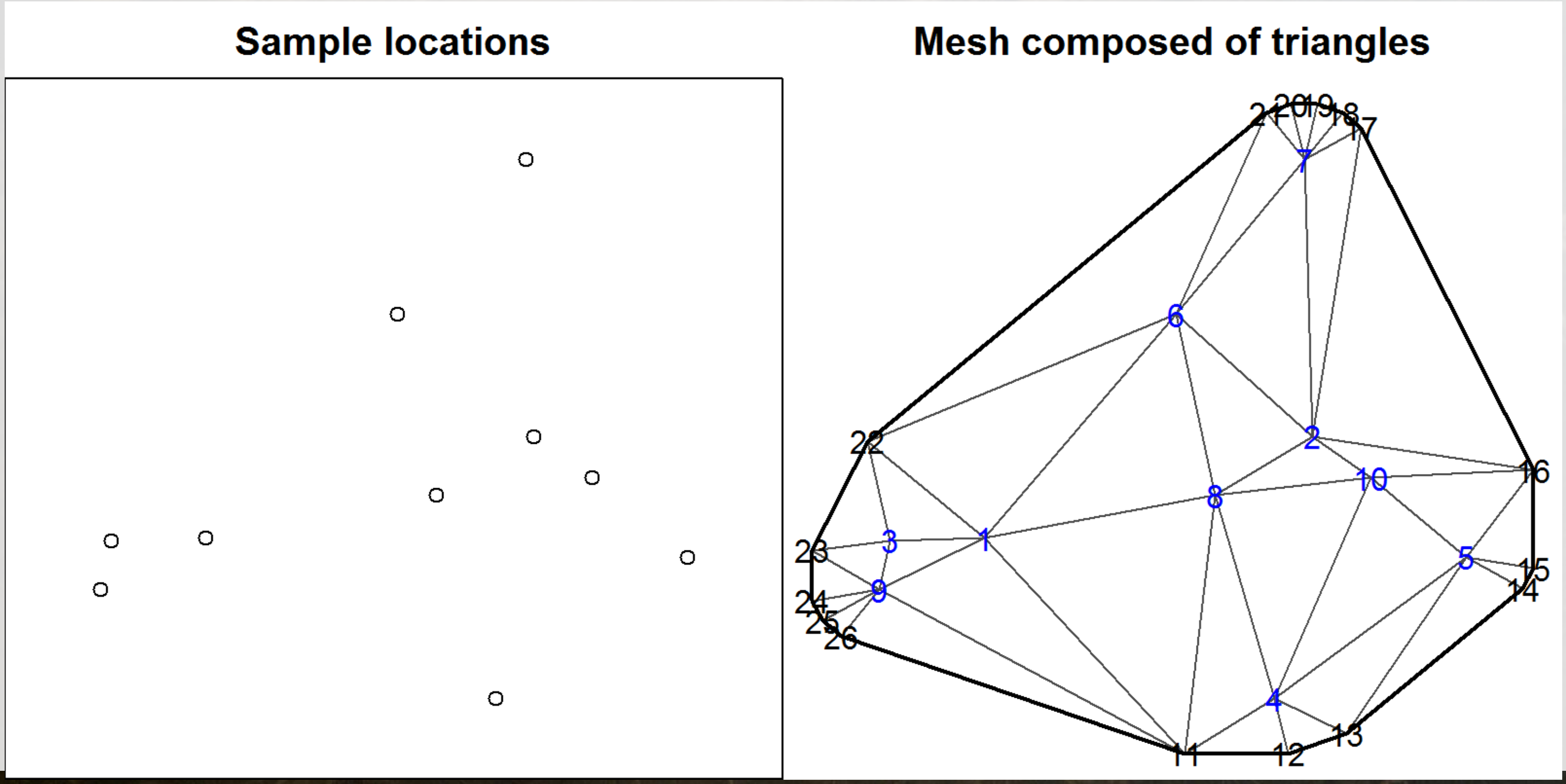


# Spatial Correlation

## Stochastic partial differential equation (SPDE)

- Separable for locations in 2D

Lindgren, Rue, Lindström. 2011. *J R Stat Soc Ser B Stat Methodol* 73(4):423–498.



# *Spatial Correlation*

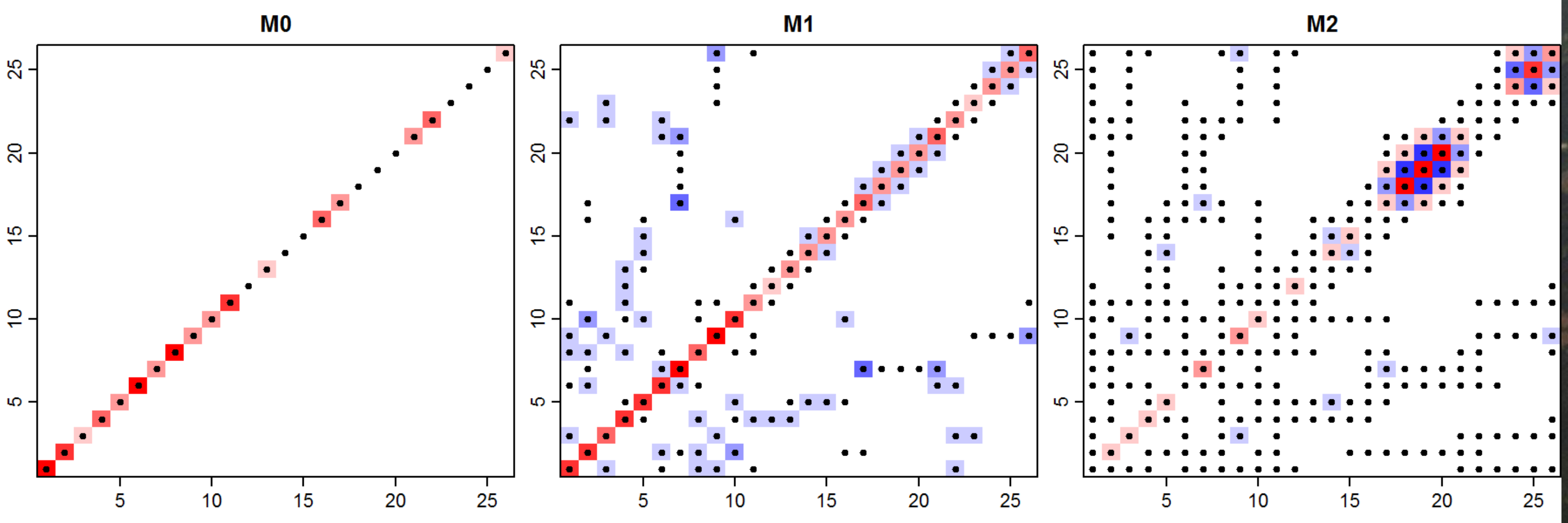
Joint distribution

$$\boldsymbol{\varepsilon} \sim MVN(0, \boldsymbol{\Sigma})$$

Which can reduce to a linear form:

$$\boldsymbol{\Sigma}^{-1} = \kappa^4 \mathbf{M}_0 + 2\kappa^2 \mathbf{M}_1 + \mathbf{M}_2$$

$$\mathbf{M}_2 = \mathbf{M}_1 {\mathbf{M}_0}^{-1} \mathbf{M}_1$$



# Four questions

- How should we impute density in areas with little data?
- **When can we use auxiliary data to separate changes in fishery catchability and fish density?**
- How should we account for non-random selection of fishing locations?
- How should we process “biological data” in conjunction with fishery CPUE?

# Vector-autogressive spatio-temporal model (VAST)

Thorson, Fonner, Haltuch, Ono, Winker. (2017) Accounting for spatiotemporal variation and fisher targeting when estimating abundance from multispecies fishery data. *Canadian Journal of Fisheries and Aquatic Sciences* **74**, 1794–1807.

- Delta-model for observations
  - Same as single-species model
- Spatio-temporal variation in density

$$\log(\lambda_i) = \alpha(t) + \sum_{f=1}^{n_f} L_\omega(c_i,f)\omega(s_i,f) + \sum_{f=1}^{n_f} L_\varepsilon(c_i,f)\varepsilon(s_i,f,t_i) + \sum_{f=1}^{n_f} L_\delta(c_i,f)\delta(f,v_i)$$

  - $\sum_{f=1}^{n_f} L_\omega(c_i,f)\omega(s_i,f)$ is spatial covariation
  - $\sum_{f=1}^{n_f} L_\varepsilon(c_i,f)\varepsilon(s_i,f,t_i)$ is spatio-temporal covariation
  - $\alpha_t$ is the intercept for each year
  - Where $\omega(f)$ and $\varepsilon(f,t)$ follow a spatial distribution with variance of one
  - $L_\omega$, $L_\varepsilon$, and $L_\delta$ are loadings matrices
- Used to predict total density

$$\hat{d}(s,c,t) = \hat{\gamma}(s,c,t) \times \hat{\lambda}(s,c,t)$$

# Fishery-dependent index standardization

- Construct indices from fishery catch rates

$$\mathbb{E}(B_c) = F_c D_c Q_c$$

  - Where
    - $B_c$ is catch for each species $c$
    - $Q_c$ is catchability
    - $F_c$ is fishing effort
    - $D_c$ is density

**3rd term in VAST catch equation**

**Goal**: Use multispecies data to "account" for fisher targeting (unexplained variation in catch-rates at a given location, caused by catchability differences)

# Joint species distribution models

## Decompose catch rates

$$\mathbb{E}(C_p) = Q_p \times F_p \times D_p$$

1. Density includes spatial variation and measured habitat variables

$$\log(D_p) = \sum_{j=1}^{J} A_{p,j}\psi_j(s,t) + \sum_{l=1}^{L} \gamma_{p,l}(t)x_l(s,t)$$

2. Fishing effort includes covariation in targeting

$$\log(F_p) = \sum_{k=1}^{K} B_{p,k}\varepsilon_k(i)$$

3. Catchability includes measured variables (i.e., GPS, plotters, vessel ID, etc.)

$$\log(Q_p) = \sum_{l=1}^{M} \nu_{p,m}y_m(i)$$

# Joint species distribution models

## Decomposing variation

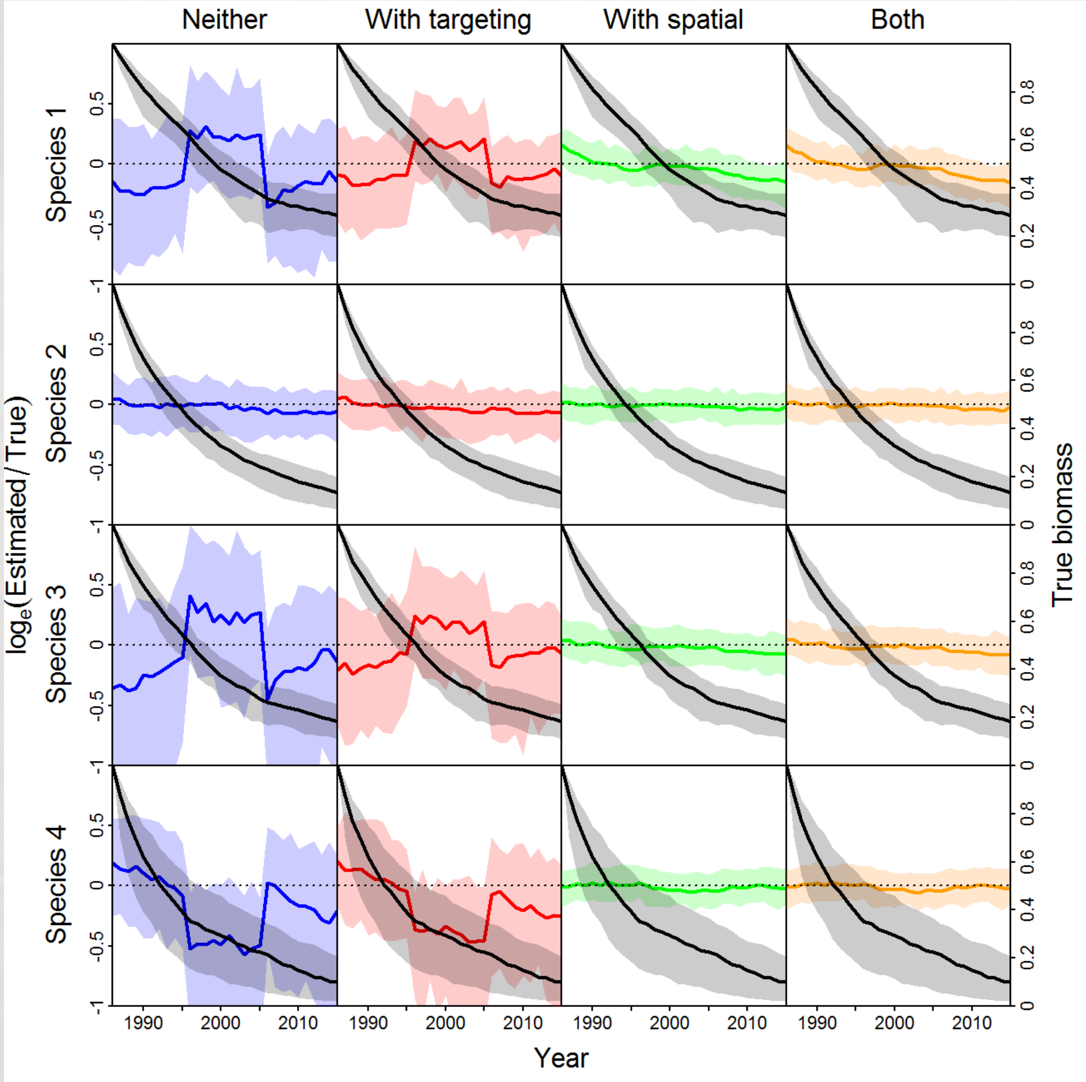
1. Spatial variation in density
   - Measurable during index standardization
2. Variation in fishing tactics
   - Not directly observed

| | Mechanisms | Model Treatment |
|---|---|---|
| **Spatial adjustments** | Initial location choice based on expected profit<br><br>Spatio-temporal adjustments in fishing location related to changes in relative ex-vessel prices of species, input costs, and regulations over time<br><br>Changes in fishing location due to new information obtained from prior fishing (e.g., avoiding areas with low catch rates) | $\mathrm{Cov}(D_p) = \mathbf{A}\mathbf{A}^T$ |
| **Tactics** | Fine-scale spatial adjustments to seek a more favorable species composition and higher catch rates once catch is observed at initial location<br><br>Changes in timing of fishing activity (e.g., daytime, nighttime, crepuscular)<br><br>Changes in fishing operations, e.g., bearing and speed<br><br>Changes in fishing gear (e.g., bait type, hook type, mesh size) | $\mathrm{Cov}(F_p) = \mathbf{B}\mathbf{B}^T$ |

# Vector-autogressive spatio-temporal model

## Simulation testing

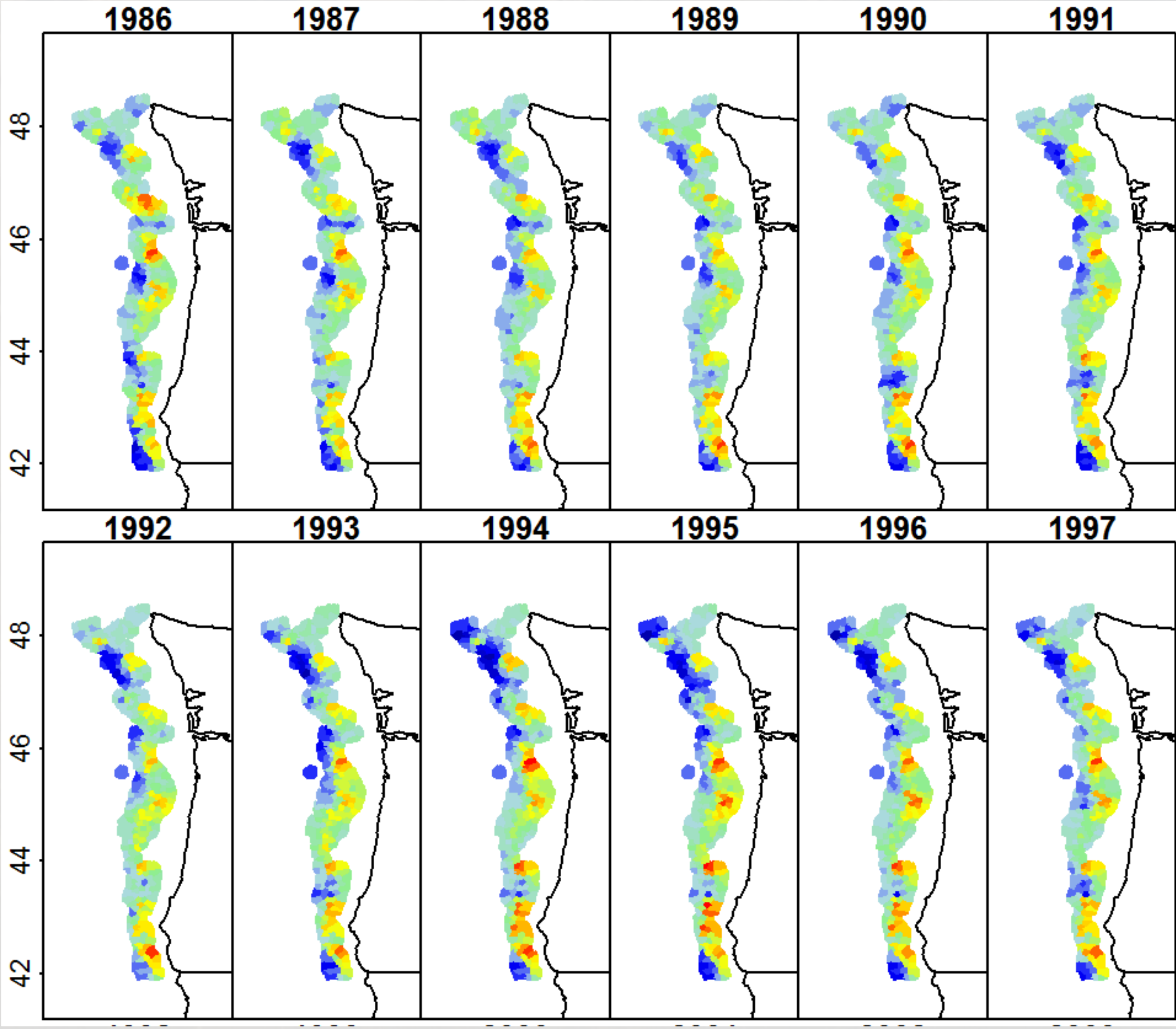
- Used simulator that was independently built
  - Generate catches for four species
- 2x2 factorial cross of four estimation models
  - With/without spatial variation
  - With/without residual targetting



# Case study: Petrale sole winter fishery

## Results

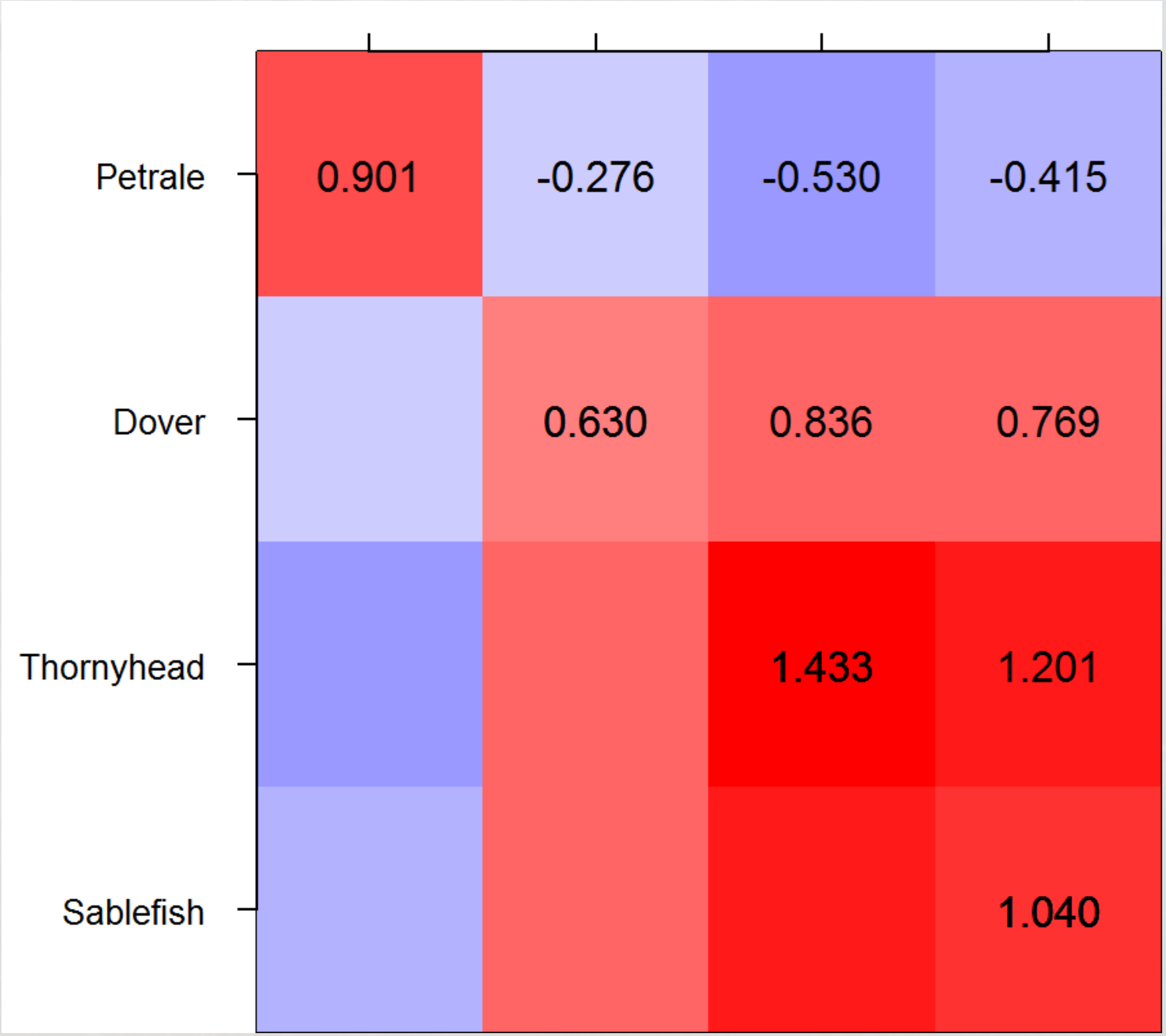
- Fit to data for Petrale, dover, sablefish, and thornyheads
- Account for targeting via residual correlations



# Covariance in catchability

- $\mathrm{Cov}(Q_c) = \boldsymbol{L}_{\boldsymbol{\delta}} \boldsymbol{L}_{\boldsymbol{\delta}}^{\boldsymbol{T}}$
- Dover, Thornyhead, Sablefish are caught together
- Winter petrale fishery is "clean"

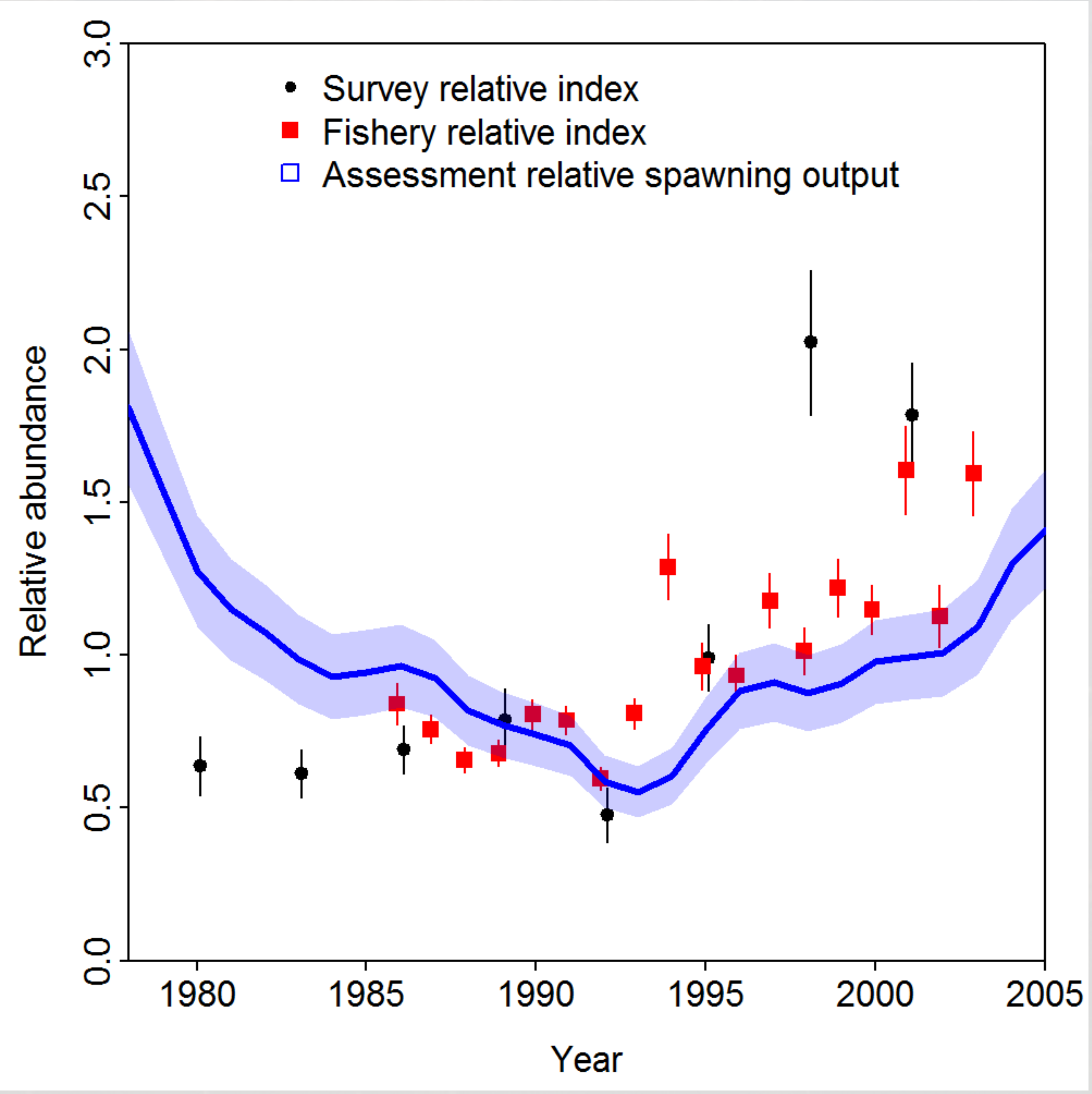
## Correlation in catchability in VAST



# Petrale sole index of abundance

**Index is plausible:**

- Matches survey index
- Timing of recovery consistent with assessment model



# Vector-autogressive spatio-temporal model

## Conclusions re: VAST

1. Can fit indices using multi-species catch-rate data
2. Residual variation in catch rates at a given location is caused by differences in catchability
   - Covaries among species…
   - … therefore catch composition is informative about catchability for a given species
3. Works well in simulation experiment
4. Provides reasonable index for Winter Petrale fishery off OR/WA
   - Corroborated by stock assessment and survey index
5. Uses similar techniques as single-species survey indices
6. Uses Travis-CI to continuously check that VAST gives identical answers to SpatialDeltaGLMM for single-species indices

# Four questions

- How should we impute density in areas with little data?
- When can we use auxiliary data to separate changes in fishery catchability and fish density?
- **How should we account for non-random selection of fishing locations?**
- How should we process "biological data" in conjunction with fishery CPUE?

# Preferential sampling

## Question

*How should we analyze data where the "design" is not independent of the "response"?*

Conn, Thorson, Johnson. (2017) Confronting preferential sampling when analysing population distributions: diagnosis and model-based triage. *Methods in Ecology and Evolution* 8, 1535–1546.

## Approach

- Simulation experiment
  - Shows sensitivity to preferential sampling
- Case study application
  - Show potential pitfalls of model-based approach

# Preferential sampling

**Definition**

- Population density $\mathcal{D}$
  - Unknown abundance in vicinity of location $s$
- Sampling intensity $\mathcal{P}$
  - Probability $\mathcal{P}(s)$ that data location $s$ will be available
- Covariates $X$
  - Could affect either density $\mathcal{D}$ or Sampling intensity $\mathcal{P}$

Preferential sampling occurs if and only if:

$$[\mathcal{D}, \mathcal{P} | X] \neq [\mathcal{D} \, | X][\mathcal{P} | X]$$

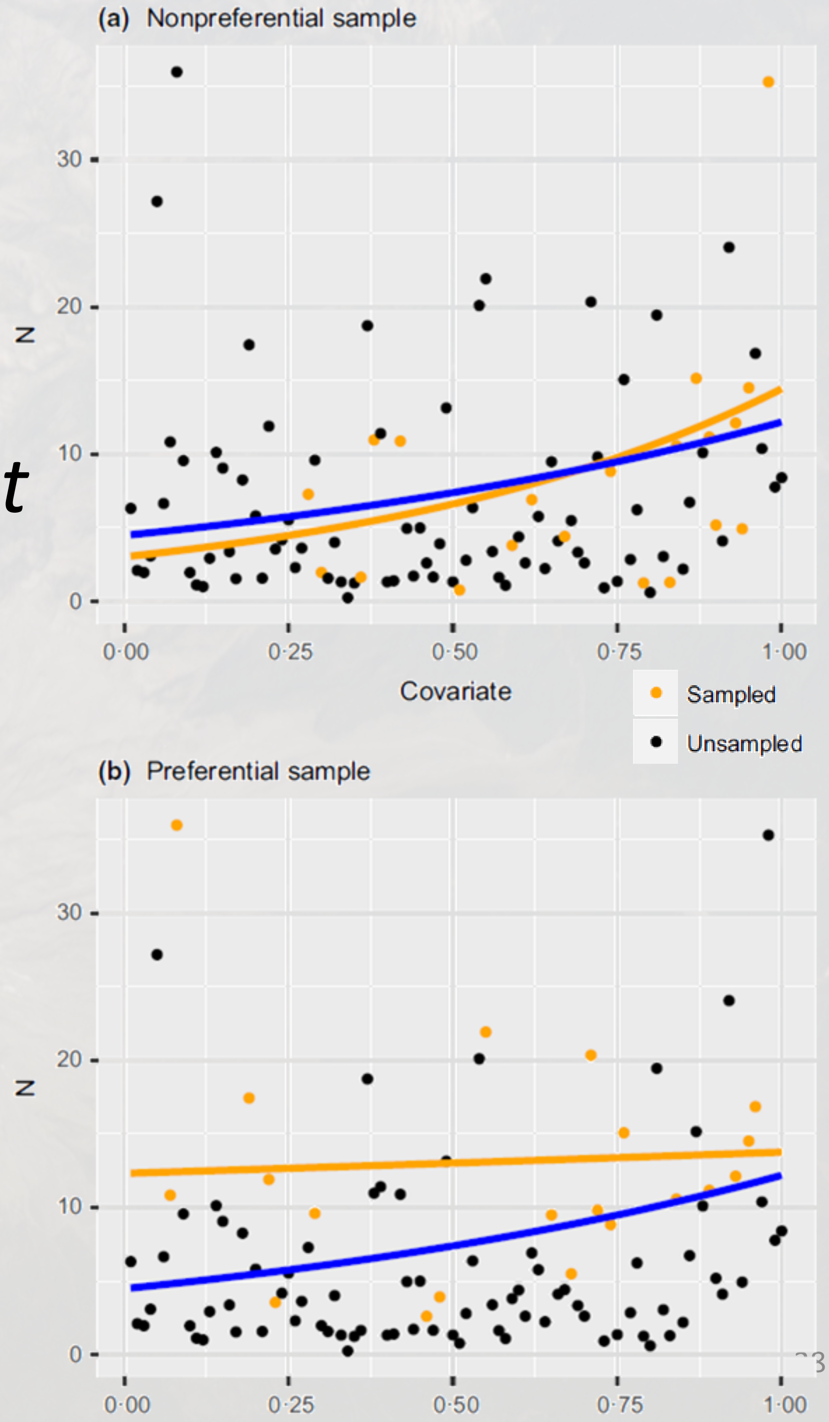
# Preferential sampling

## Problem

*Causes bias because no samples in low-density habitat*

## Solution

- Jointly model sampling intensity and density
- Use estimated density to extrapolate density into areas with no data



# Preferential sampling

**Simulation experiment**:

- Simulate density

$$\log\big(\lambda(s)\big) = \beta_0 + \mathbf{x}(s)\boldsymbol{\beta} + \omega(s)$$

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

- Simulate inclusion probability

$$\mathrm{logit}\big(\nu(s)\big) = \beta_0^* + \mathbf{x}(s)\boldsymbol{\beta}^* + \psi(s) + b\omega(s)$$

$$\boldsymbol{\psi} \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}^*)$$

- Simulate location of data
  - Draw 50 locations **s** from $\nu(s)$
- Simulate data

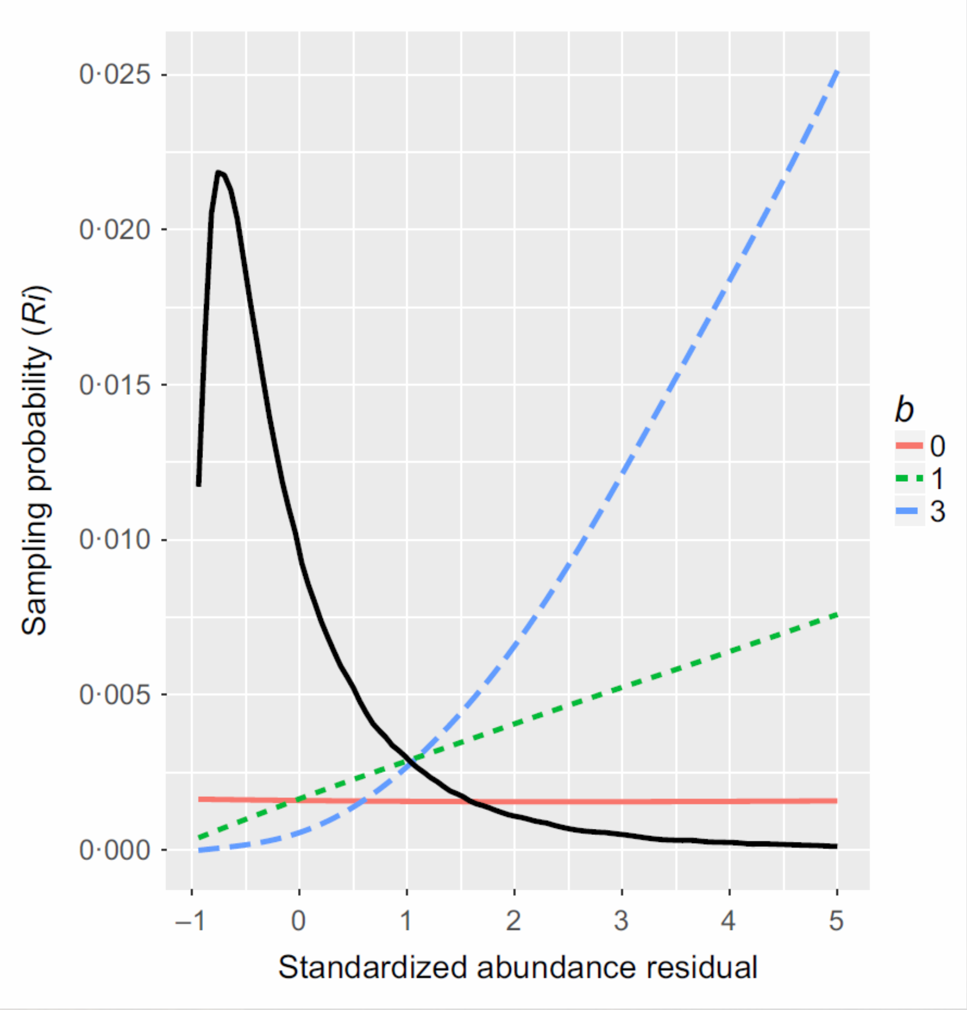
$$c(s) \sim Poisson\big(\lambda(s)a(s)\big)$$

# Preferential sampling

**Simulation experiment**:

- Uses areal formulation
  - Small differences in notation but otherwise similar
- Three scenarios
  - Not preferential: $b = 0$
  - Weakly preferential: $b = 1$
  - Strongly preferential: $b = 3$
- 500 replicates per scenario
  - Uses “epsilon bias-correction”
- Evaluates error in total abundance

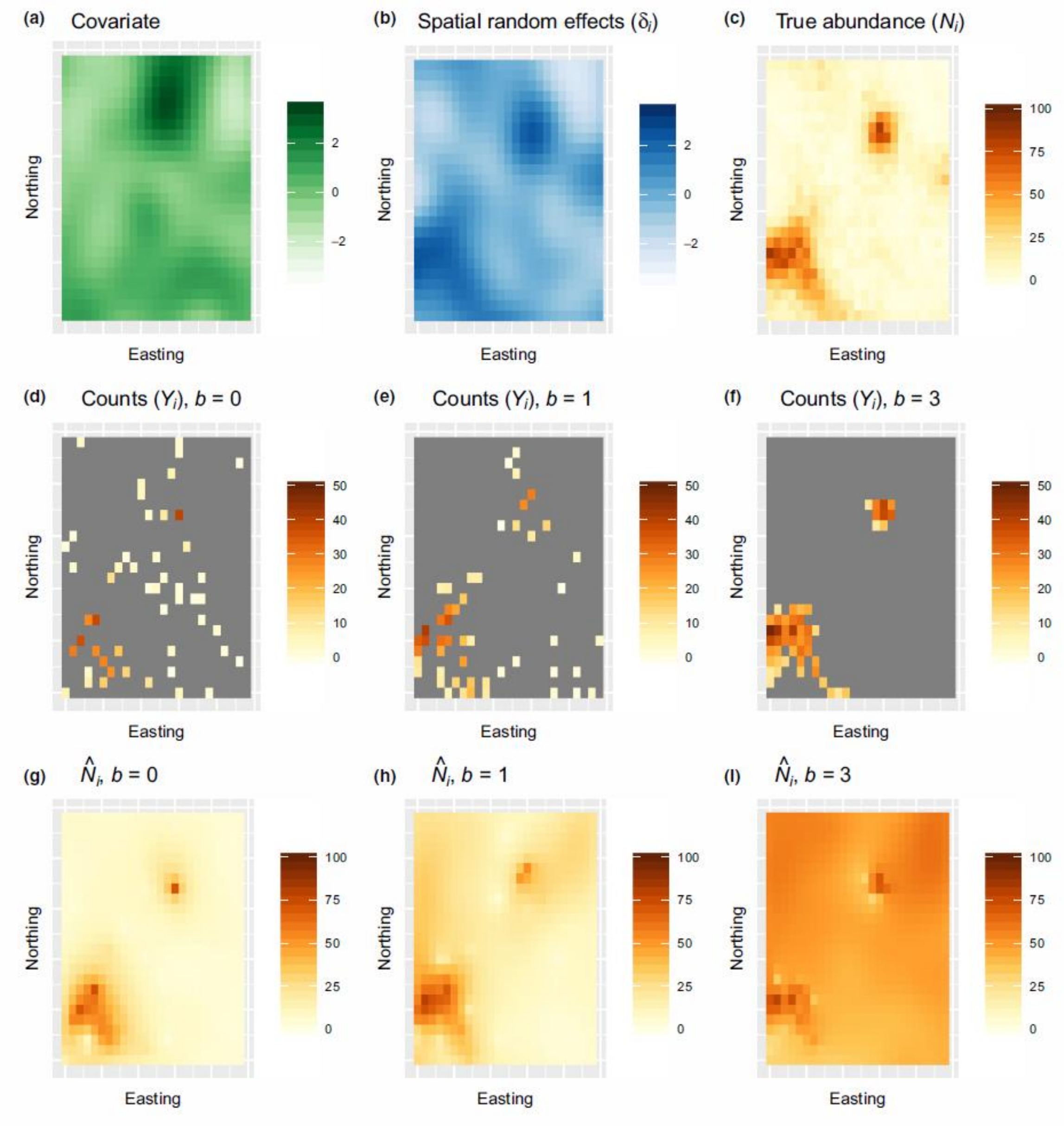
$$N_{total} = \sum \lambda(s) a(s)$$



# Preferential sampling

**Simulation experiment**:

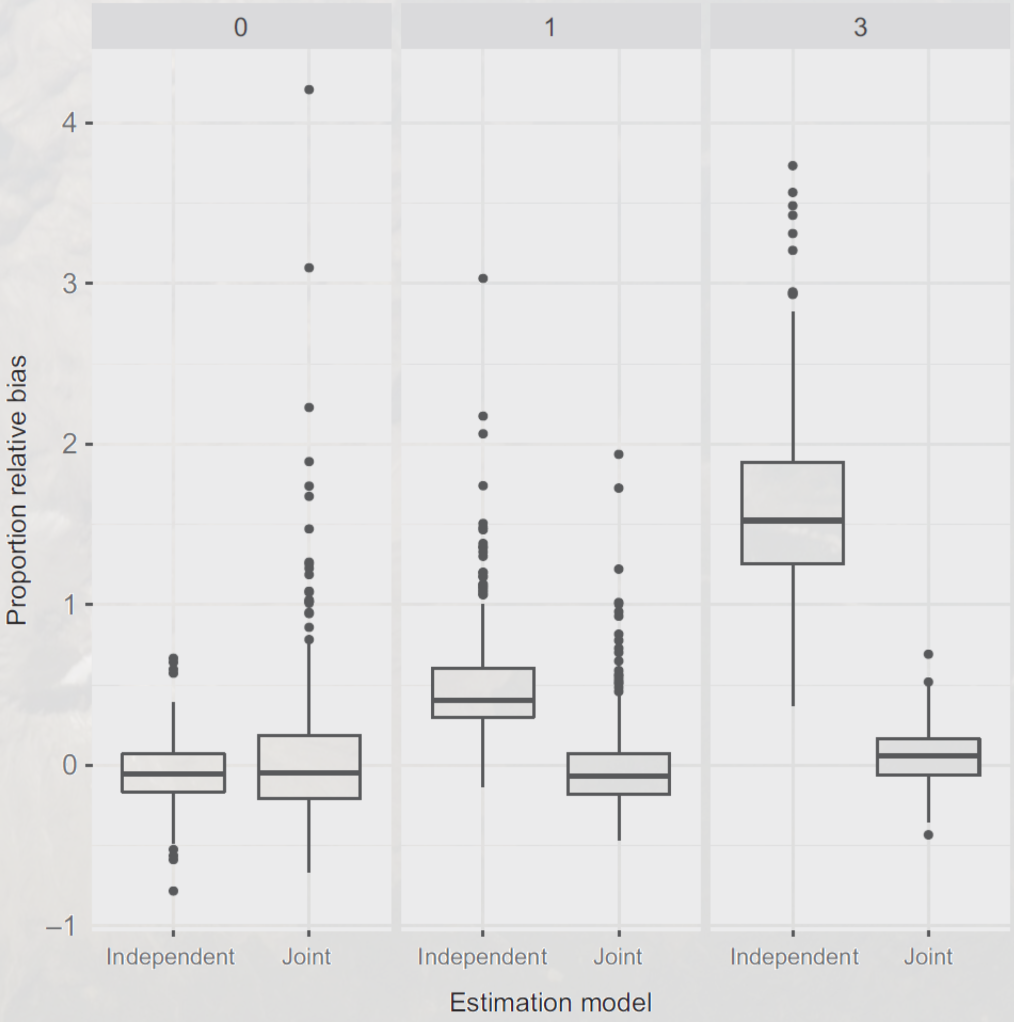
- Shows non-independence
  - More samples in high-density areas when $b = 3$
- Shows potential bias
  - Over-estimate density in unsampled areas when $b = 3$



# Preferential sampling

**Simulation experiment**:

- Biased when ignoring preferential sampling
  - 50% bias when $b = 1$
  - 150% bias when $b = 3$
- Increased error when estimating $b$ when $b = 0$



# Preferential sampling

**Case study**

- Model selection differs for different criteria

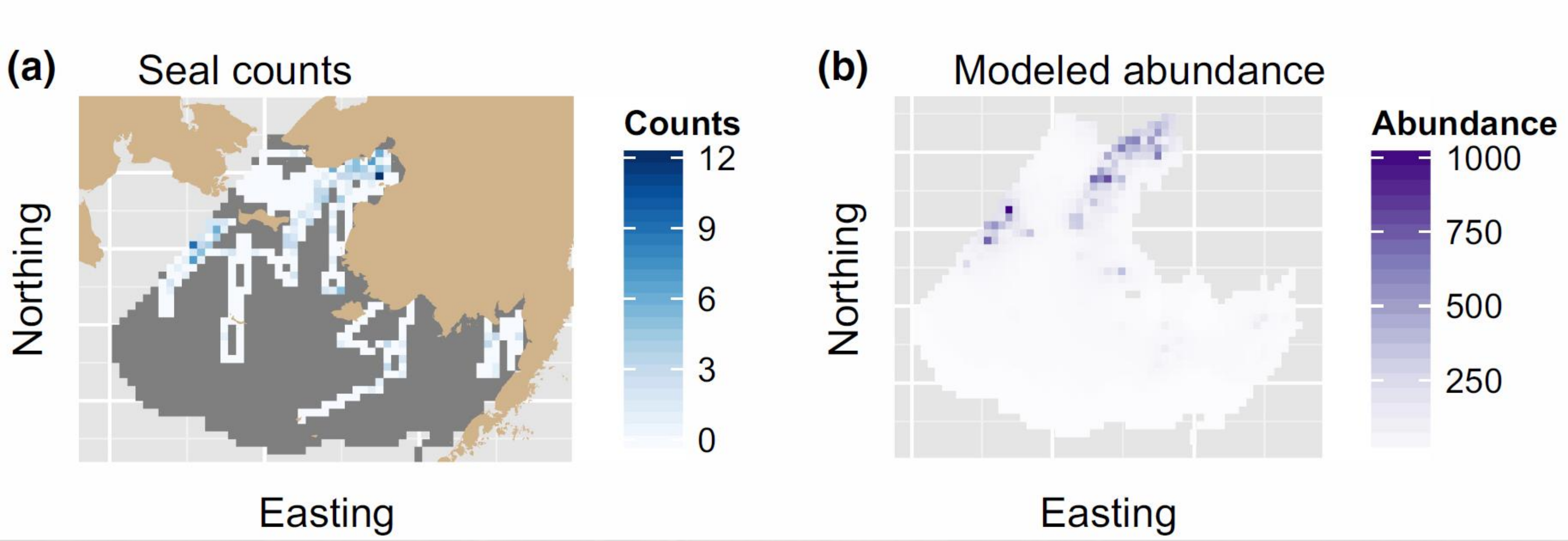
| **Include covariate** | **Include pref. sampling** | **Number of params.** | **Cross-validation error** | $\Delta AIC$ | $\hat{N}$ | $\widehat{SE}(N)$ |
|---|---|---|---|---|---|---|
| No | No | 5 | 87.2 | 99.1 | 70,738 | 6,988 |
| No | Yes | 6 | 116.2 | 1.9 | 43,232 | 2,778 |
| Yes | No | 12 | 88.3 | 103.0 | 66,989 | 20,374 |
| Yes | Yes | 13 | 105.3 | 0.0 | 40,656 | 3,664 |

- AIC selects covariate + preferential sampling
- Cross-validation selects neither one

# Preferential sampling

**Case study**

- Showing results for model without preferential sampling




- BUT: results highly sensitive to model decisions

# Preferential sampling

**Synopsis**

- Preferential sampling causes bias due to poor extrapolation in unsampled areas
- Joint models can mitigate bias
  - True only if the model is correctly specified
- Results are sensitive to model specification
  - Selected model may differ among criteria
- It is possible to implement using package VAST
  - Treat sampling intensity as a $2^{nd}$ “species”
  - Multivariate dimension reduction could be useful given data from many different sources

# Four questions

- How should we impute density in areas with little data?
- When can we use auxiliary data to separate changes in fishery catchability and fish density?
- How should we account for non-random selection of fishing locations?
- **How should we process “biological data” in conjunction with fishery CPUE?**

# Spatio-temporal comp-expansion

## Question

How to expand subsamples from survey tows?

## Approach

- Expand subsamples to biomass for each tow
- Analyze catch for each category using multivariate spatio-temporal model
- Process variance estimates to calculate “input sample size”
  - Input sample size ≡ multinomial sample size with same variance

# Spatio-temporal comp-expansion

## Details

1. Fit delta-model to numbers $n_c(i)$ for each category $c$

$$\Pr(n_c(i) = B) = \begin{cases} 1 - p_c(i) & \text{if } B = 0 \\ p_c(i) \times Lognormal(B | r_c(i), \sigma_m^2(c)) & \text{if } B > 0 \end{cases}$$

2. Predictors in delta-model include spatio-temporal variation

$$logit\big(p_c(i)\big) = \beta_p(c, t_i) + \sigma_{\omega p}(c)\omega_p(c, s_i) + \sigma_{\varepsilon p}(c)\varepsilon_p(c, s_i, t_i)$$

$$log(r_c(i)) = \log(a_i) + \beta_r(c, t_i) + \sigma_{\omega r}(c)\omega_r(c, s_i) + \sigma_{\varepsilon p}(c)\varepsilon_r(c, s_i, t_i)$$

3. Assemble index by category

$$\hat{d}_c(s, t) = \hat{p}_c(s, t) \times \hat{r}_c(s, t)$$

$$\hat{I}_c(t) = \sum_{s=1}^{n_s} \big(a(s) \times \hat{d}_c(s, t)\big)$$

# Spatio-temporal comp-expansion

## Details

4. Calculate standard error for proportions

$$\widehat{\mathrm{SE}}\left[\hat{P}_c(t)\right]^2 \approx \frac{\hat{I}_c(t)}{\hat{I}(t)}\left[\frac{\widehat{\mathrm{SE}}\left[\hat{I}_c(t)\right]^2}{\hat{I}_c(t)} + \frac{\sum_{c=1}^{n_c}\widehat{\mathrm{SE}}\left[\hat{I}_c(t)\right]^2}{\hat{I}(t)}\right]$$

5. Calculate "input sample size" $\hat{\tau}(t)$

$$\hat{\tau}(t) = \mathrm{Median}_c\left[\frac{\hat{P}_c(t)\left(1-\hat{P}_c(t)\right)}{\widehat{\mathrm{SE}}\left[\hat{P}_c(t)\right]^2}\right]$$

# Spatio-temporal comp-expansion

## Simulation experiment

- Age-structured spatio-temporal "Operating model" (OM)
  - Abundance at age

$$N_a(s,t) = \begin{cases} \exp\big(\beta_N + \omega_N(s) + \varepsilon_N(s,t)\big) \times \exp(-Za) & \text{if } t = 1 \text{ or a} = 1 \\ N_{a-1}(s,t-1) \times \exp(-Z) & \text{if } t > 1, a > 1 \end{cases}$$

  - Biomass at age

$$W_a(s,t) = w_\alpha(L_\infty \exp(-Ka))^{w_\beta} \times \exp\big(\omega_W(s) + \varepsilon_W(s,t)\big)$$

  - Simulated sampling

$$p_i(a) = 1 - \exp\big(-a_i S_a N_a(s_i,t_i)\big)$$

$$r_i(a) = \frac{a_i S_a N_a(s_i,t_i)}{p_i(a)} \times W_a(s_i,t_i)$$

  - Simulated "true" proportion at age

$$P_c(t) = \frac{\sum_{s=1}^{n_s}(a(s) \times N_a(s,t) \times W_a(s,t))}{\sum_{a=1}^{n_a}\sum_{s=1}^{n_s}(a(s) \times N_a(s,t) \times W_a(s,t))}$$

# Spatio-temporal comp-expansion

## Simulation experiment

- Performance criteria

1. Error

2. Confidence interval coverage

$$\chi^2(t) = \sum_{c=1}^{n_c} \hat{\tau}(t)\, \hat{P}_c(t) \log\left(\frac{\hat{P}_c(t)}{P_a(t)}\right)$$

$$Q(t) = \int_0^{\chi^2(t)} Chi.squared(n_a)$$

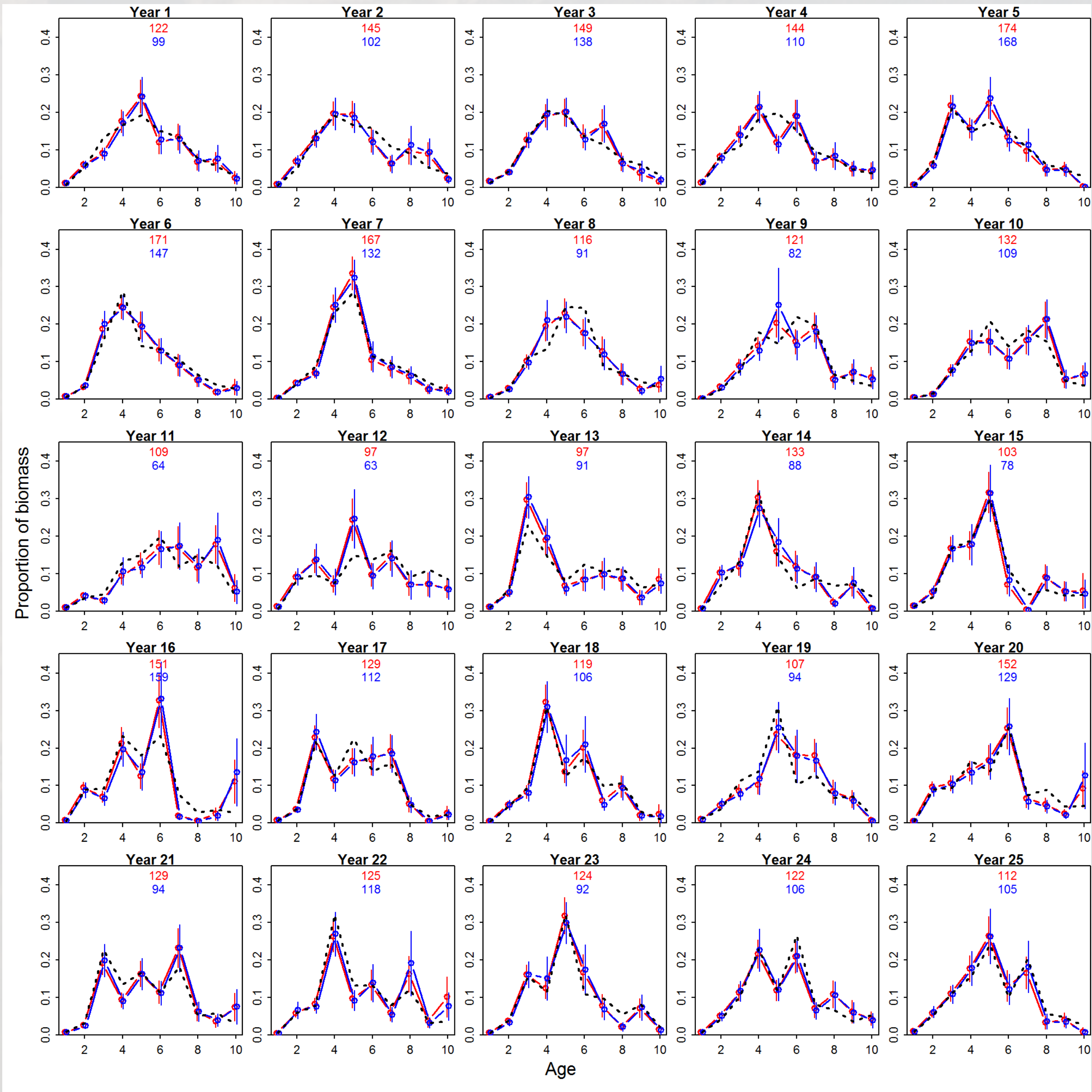
where $Q(t)$ should be uniform

# Spatio-temporal comp-expansion

Design-based Spatio-temporal True

## Simulation results

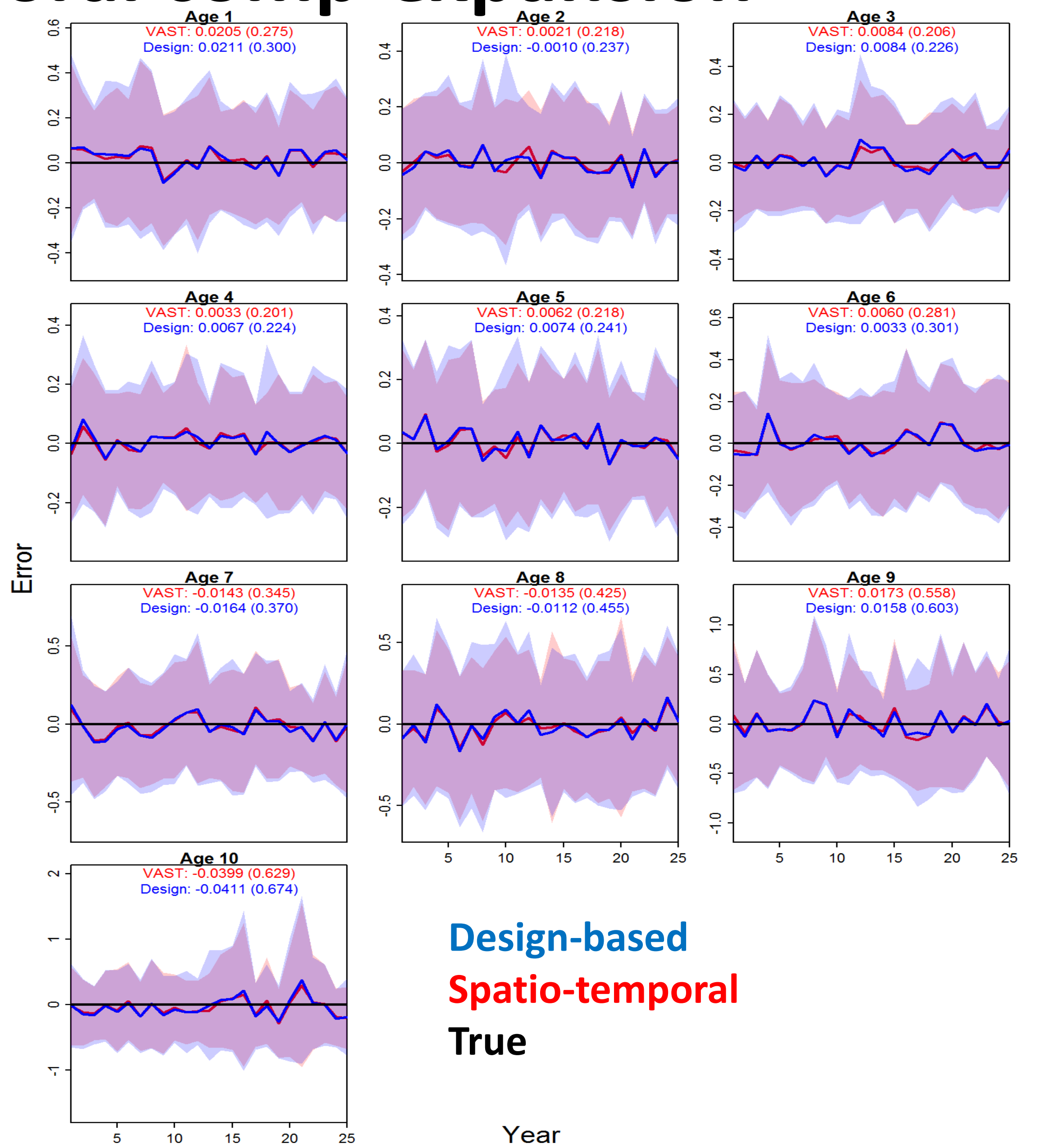
- Design and spatial provide similar results
- Can track cohorts through OM and both EMs



# Spatio-temporal comp-expansion

## Simulation results

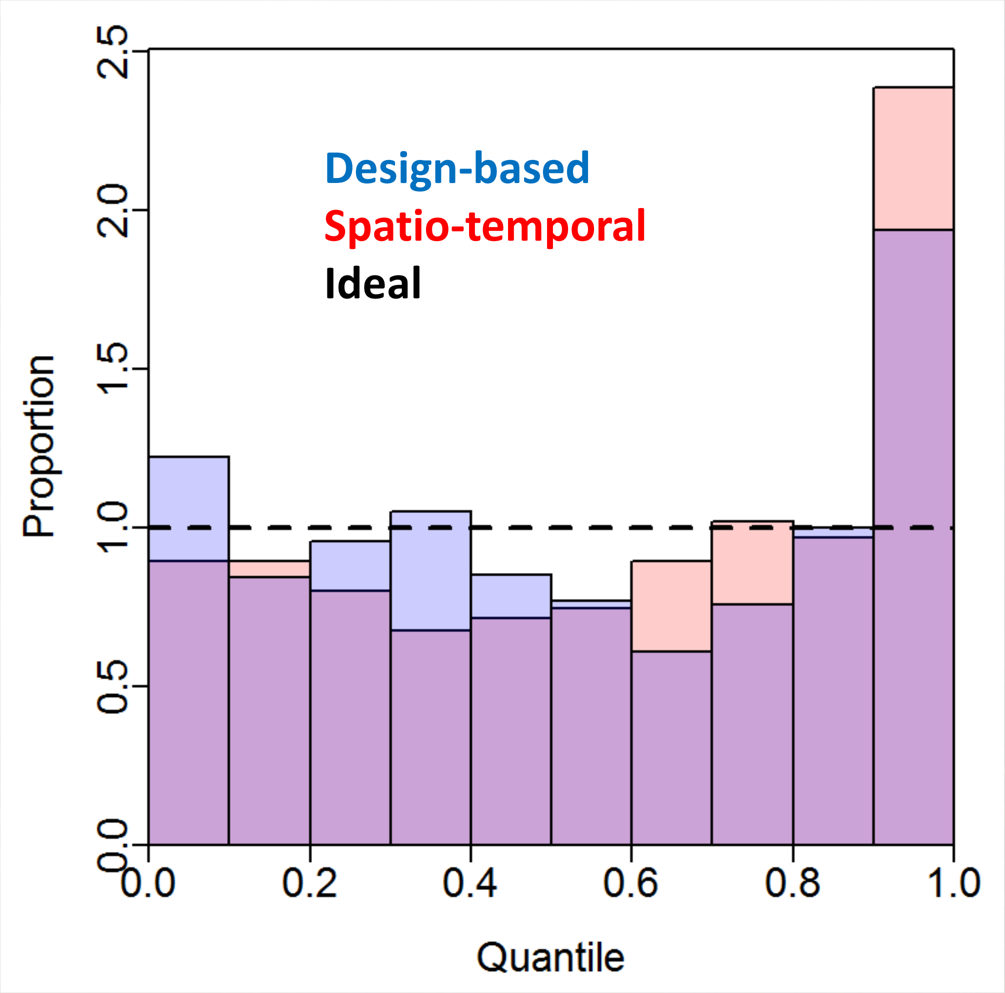
- Both are essentially unbiased (number at top of each panel)
- Spatial has 10-25% decrease in root-mean-squared error (parentheses)



# Spatio-temporal comp-expansion

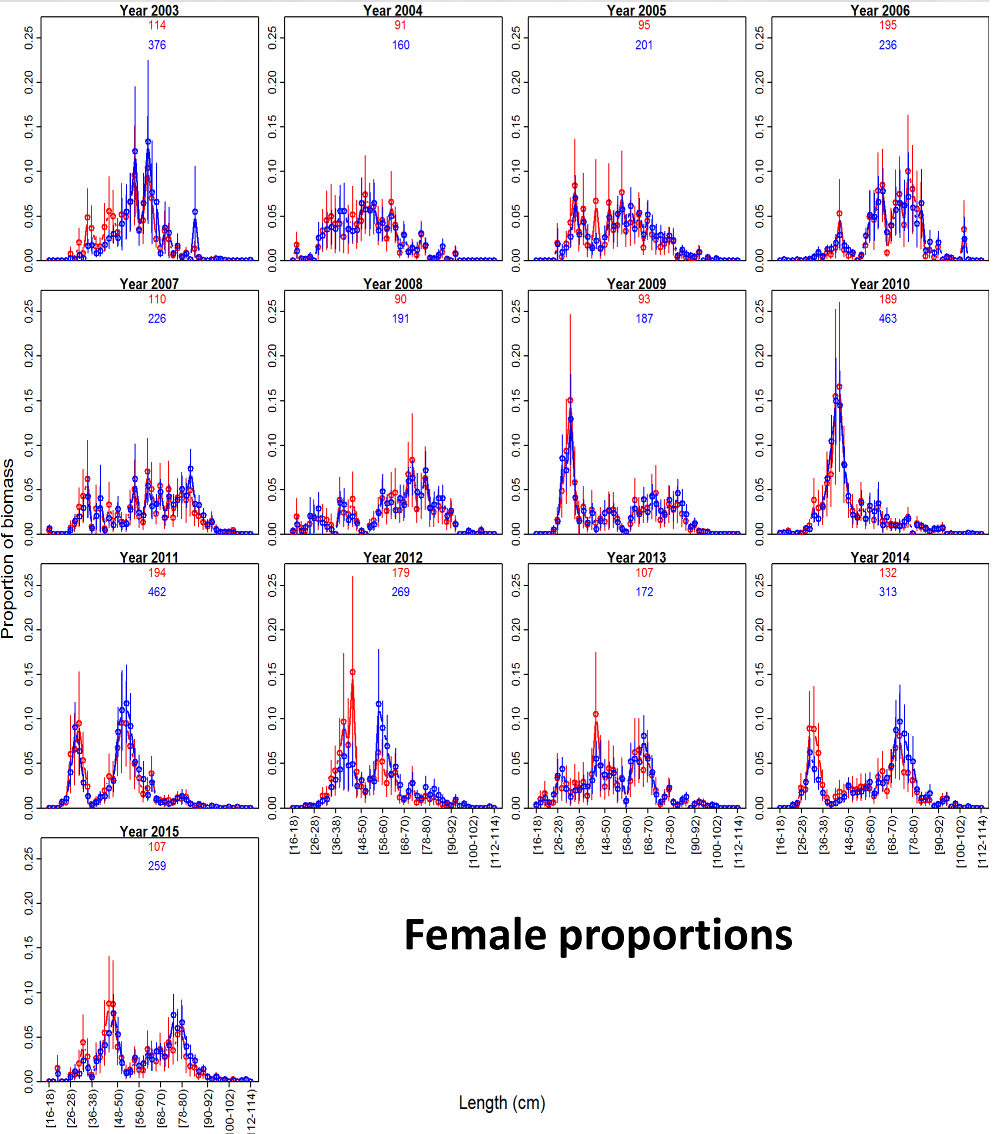
## Simulation results

- Both design and spatio-temporal have OK coverage
- Both have an excess of $Q(t) \to 1$
  - Replicates where input sample size is too small!

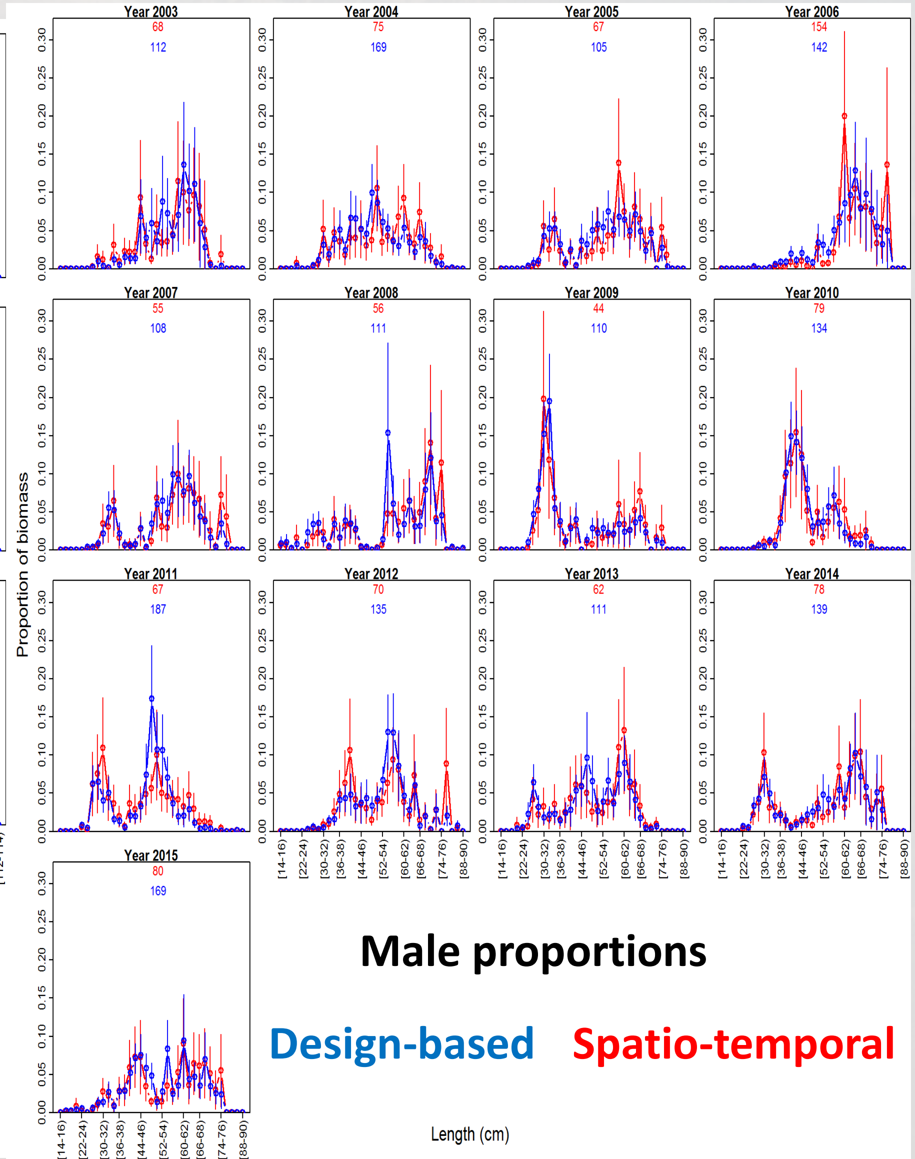


# Spatio-temporal comp-expansion

## Lingcod case-study application




**Female proportions**




**Male proportions**

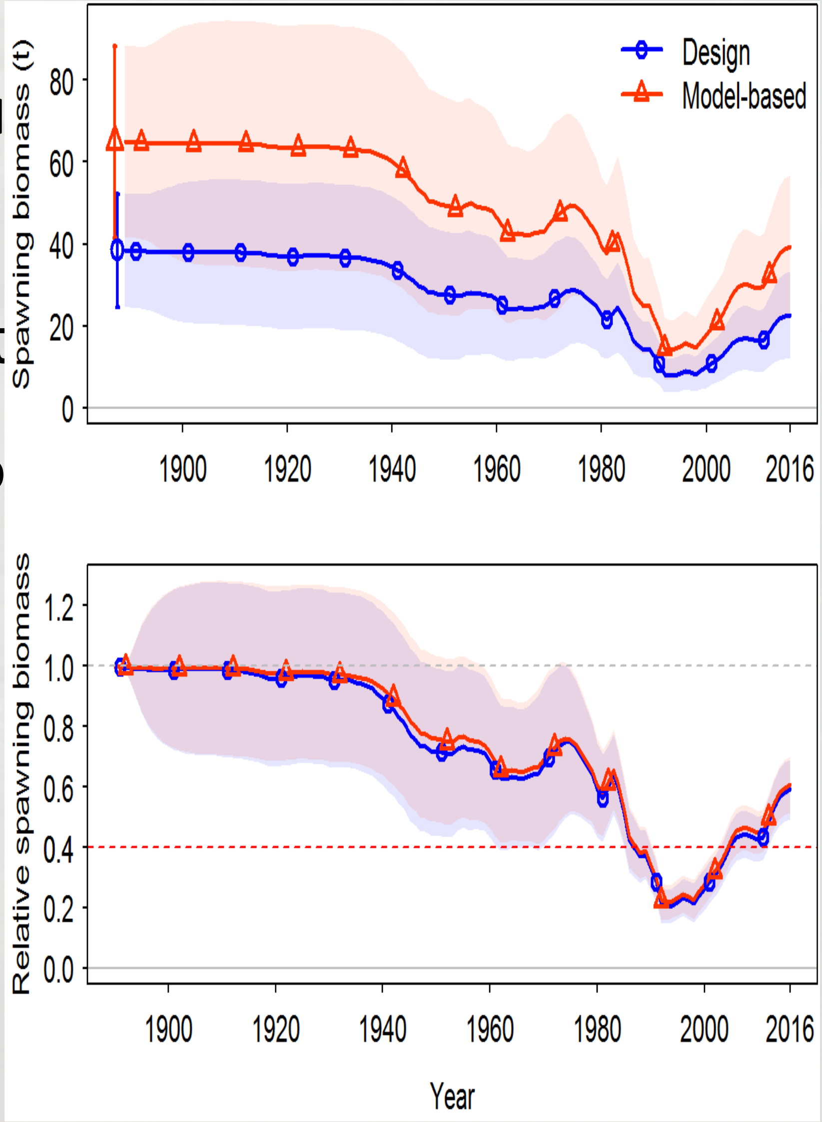
**Design-based** **Spatio-temporal**

# Spatio-temporal comp-expansion

## Lingcod case-study application

- Makes a big difference in absolu
  assessment model

- Spatio-temporal has lower input
  - Sample size doesn't seem so impo



# Spatio-temporal comp-expansion

## Conclusions

1. It is computationally feasible to do comp-expansion using spatio-temporal model
   – Can even use 2 cm bins with separate male vs. female
2. Not clear that there's a big benefit
   – Simulation showed a 25% decrease in root-mean-squared error
   – Case study showed increase in RMSE
   – Case study showed a large impact on assessment results

# Four questions

- How should we impute density in areas with little data?
- When can we use auxiliary data to separate changes in fishery catchability and fish density?
- How should we account for non-random selection of fishing locations?
- How should we process “biological data” in conjunction with fishery CPUE?

# Conclusions

## We know how to…

- Extrapolate density in unsampled areas
- Use auxiliary data to identify residual targeting
- Account for non-random availability of data
- Expand biological (age/length) data within spatio-temporal models

# Conclusions

## Next steps

- Explore applications in diverse fisheries
  - Different magnitude of missing-data problems
  - Different “information content” in multispecies data
- Scale-up to larger problems
  - Many high-seas data sets have >10,000,000 observations
  - Some regions have substantial variation at <1km resolution
- Integrate multiple data types
  - We sometimes have a mix of fishery and survey data
  - Fishery data might be presence-only, presence/absence, count, or biomass-sampling records

# Acknowledgements

## Co-authors:

- Paul Conn
- Devin Johnson
- Melissa Haltuch
- Ole Shelton
- Eric Ward
- Hans Skaug
- Kasper Kristensen
- Robby Fonner
- Kotaro Ono

## Organizers

- Mark Maunder
- Kevin Hill