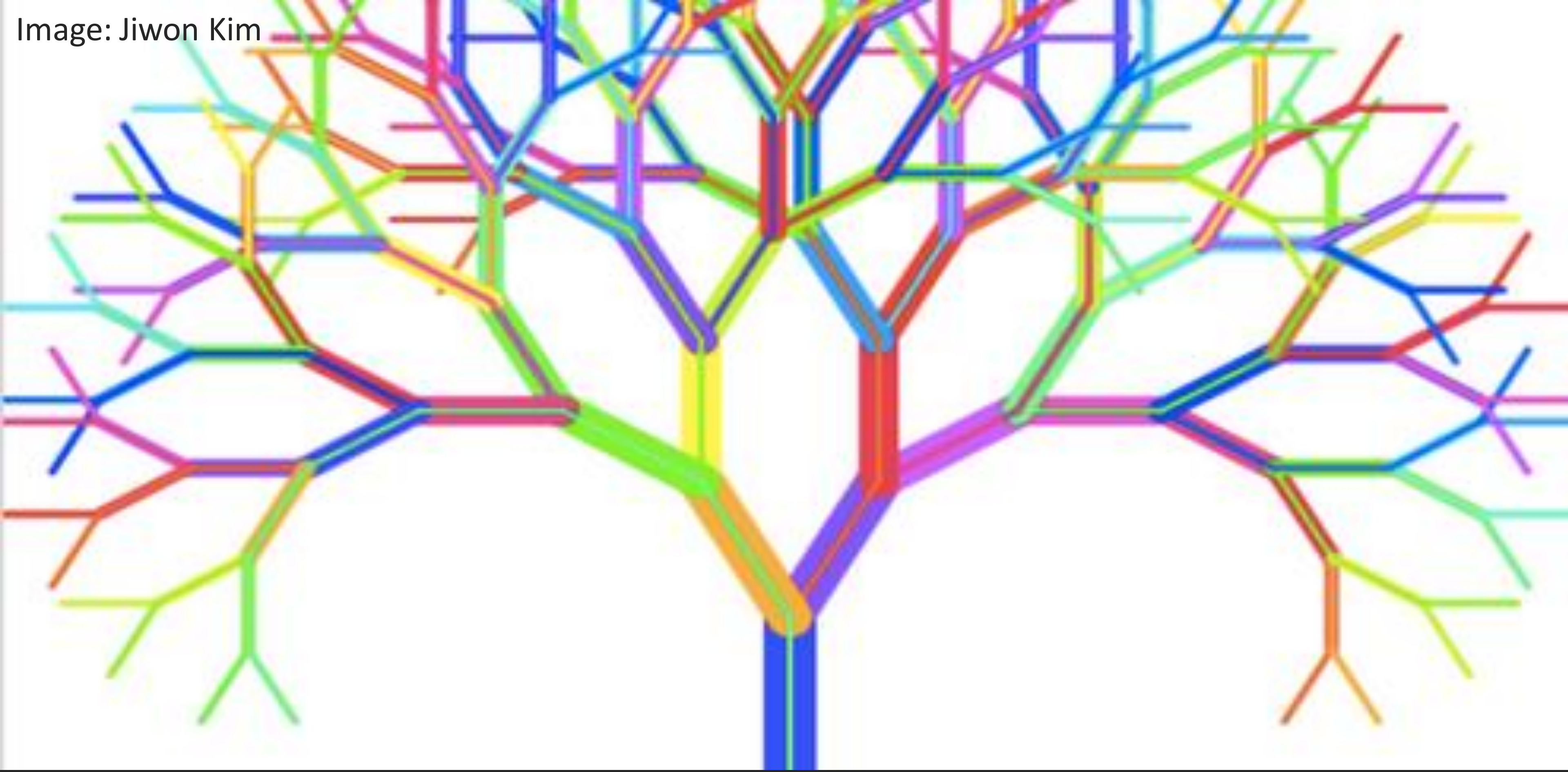


Image: Jiwon Kim

Can we use random forests for spatiotemporal CPUE modeling?

BRIAN STOCK, ERIC WARD, BRICE SEMMENS

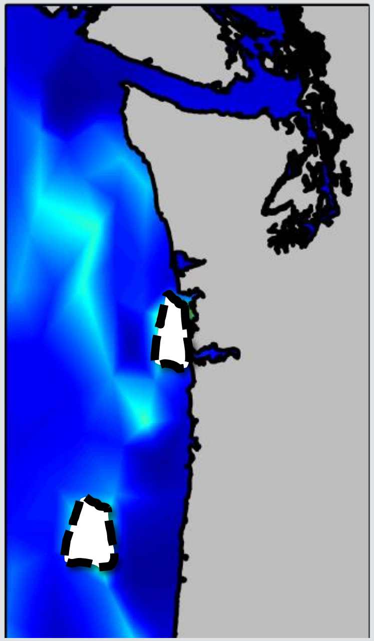
What we want (from Rick Methot)

- Fast (coding vs. runtime vs. interpretation)
- Replicable (method well-defined, get same answer)
- Robust (insensitive to distributional assumptions, outliers)
- Predictive ability (minimal errors, fill in space/time gaps)
- Covariate effects (nonlinear, interactions)
- -- Uncertainty estimates (with known properties)
- Specifiable structure (e.g. correlation through time, biology)
- Unbiased (relative vs. absolute abundance)

What we want (from Rick Methot)

- Fast
- Replicable
- Robust
- Predictive ability
- Covariate effects
- -- Uncertainty estimates
- Specifiable structure
- Unbiased

Story 1: Bycatch hotspots

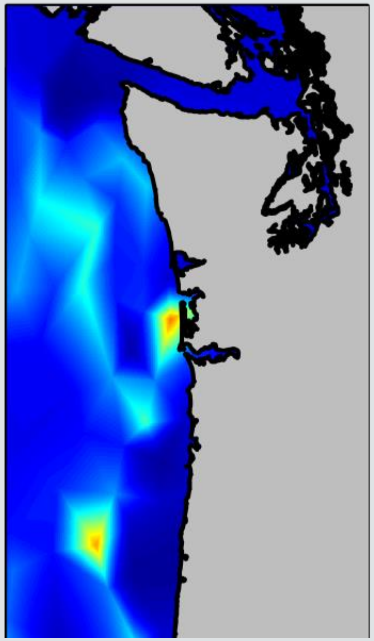


What we want (from Rick Methot)

- Fast
- Replicable
- Robust
- Predictive ability
- Covariate effects
- -- Uncertainty estimates
- Specifiable structure
- Unbiased

Story 2: Total bycatch estimation

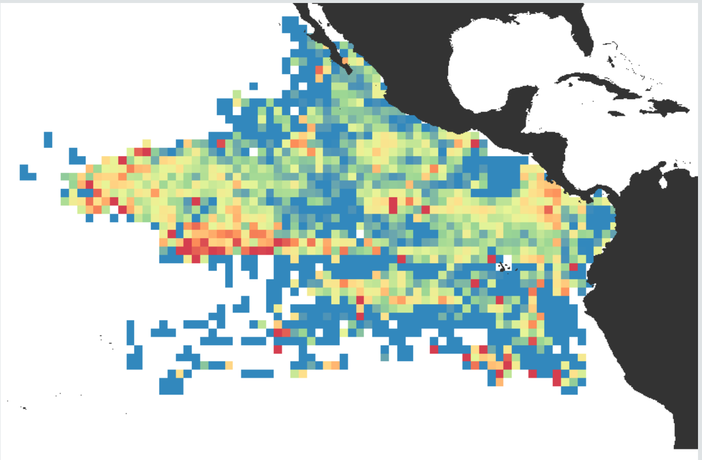
$$\sum$$



What we want (from Rick Methot)

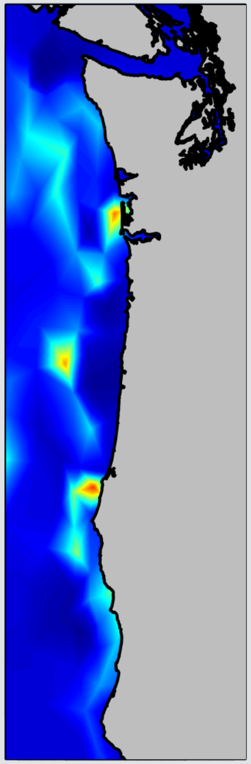
- Fast
- Replicable
- Robust
- Predictive ability
- Covariate effects
- -- Uncertainty estimates
- Specifiable structure
- Unbiased

Story 3: CPUE standardization



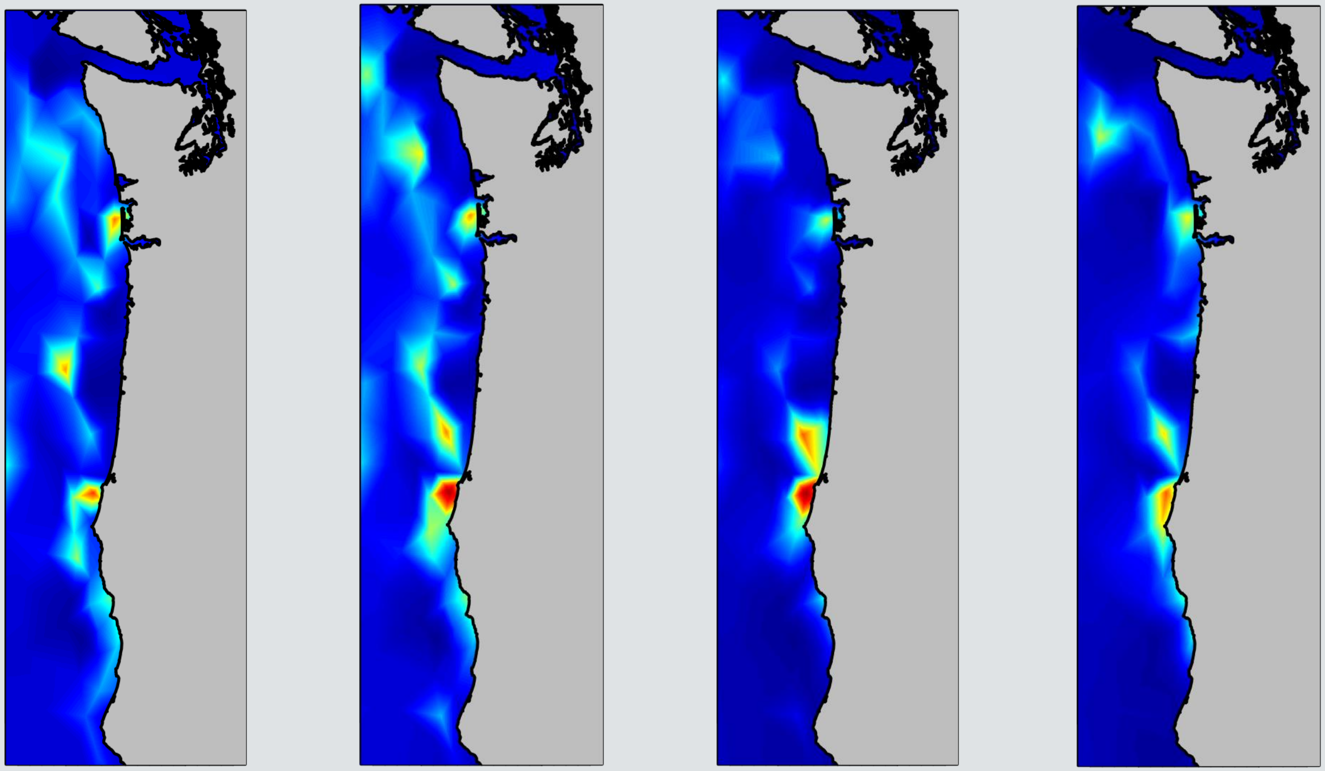
Tools for dynamic management

Need map of bycatch "risk"



Tools for dynamic management

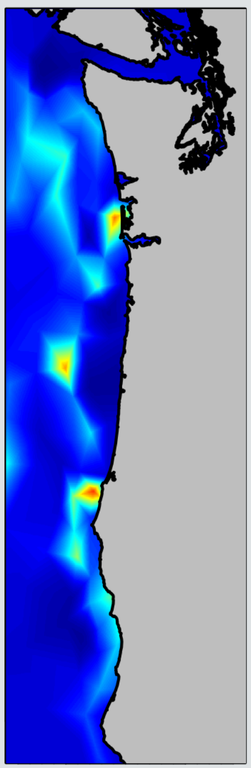
Need map of bycatch "risk"



1. Introduction

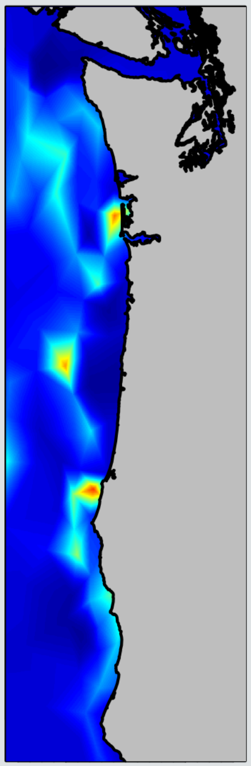
Tools for dynamic management

Need map of bycatch "risk"



- temperature
- depth
- substrate
- spatial field

Q: Which spatial model is best?



- temperature
- depth
- substrate
- spatial field

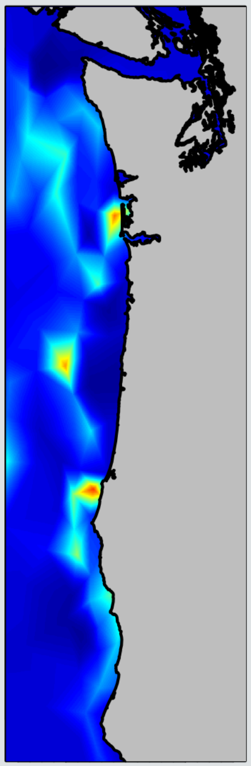
GLM

GAM

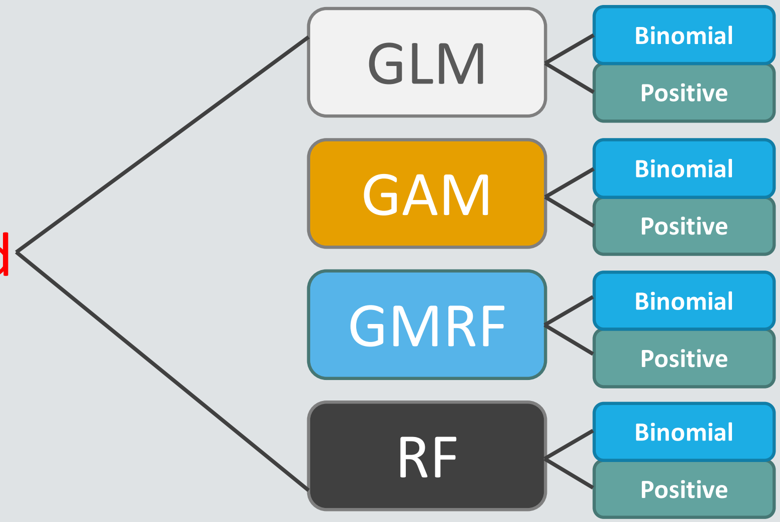
GMRF

RF

Q: Which spatial model is best?



- temperature
- depth
- substrate
- spatial field



What are these models exactly?

GLM

GAM

GMRF

RF

obs ~ environmental predictors (temp, depth, …)

$$Y_i \sim Binomial(logit^{-1}[\boldsymbol{X_i\beta}])$$

Binomial

$$Y_i \sim Gamma(e^{\boldsymbol{X_i\beta}}, \nu)$$

Positive

What are these models exactly?

GLM

GAM

GMRF

RF

obs ~ environmental predictors (temp, depth, …)

obs ~ environmental predictors + s(lat,lon)

What are these models exactly?

GLM — obs ~ environmental predictors (temp, depth, …)

GAM — obs ~ environmental predictors + s(lat,lon)

GMRF — obs ~ environmental predictors + $MVN(0, \boldsymbol{\Sigma})$

RF

What are these models exactly?

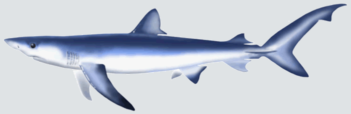
GLM	obs ~ environmental predictors (temp, depth, …)
GAM	obs ~ environmental predictors + s(lat,lon)
GMRF	obs ~ environmental predictors + $MVN(0, \boldsymbol{\Sigma})$
RF	obs ~ environmental predictors + lat + lon

Fisheries observer data

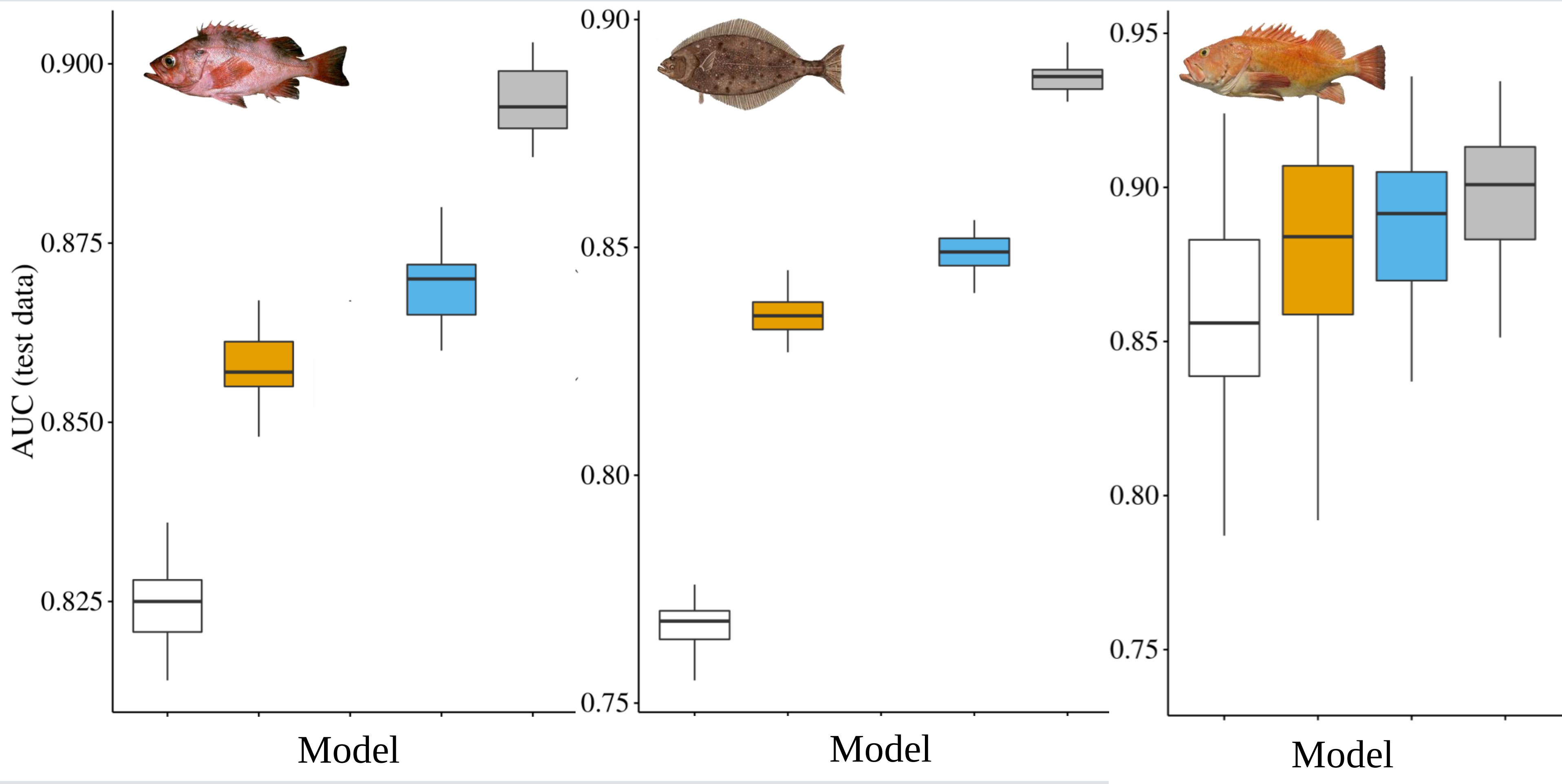
U.S. West Coast
Groundfish
Trawl



Hawaii
Swordfish
Longline



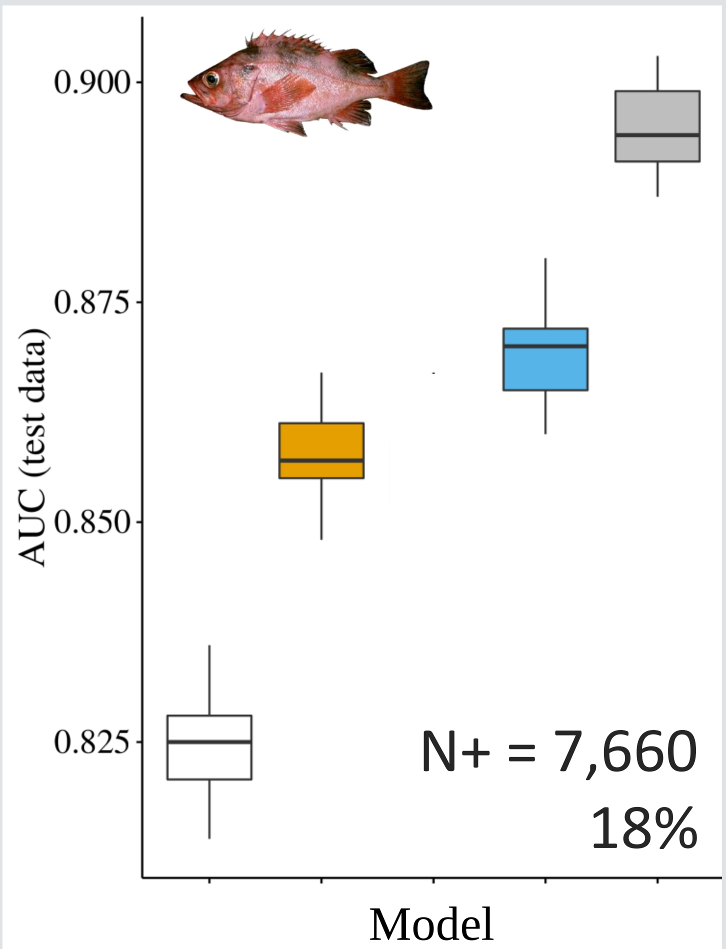
Generally: GLM < GAM < GMRF < RF

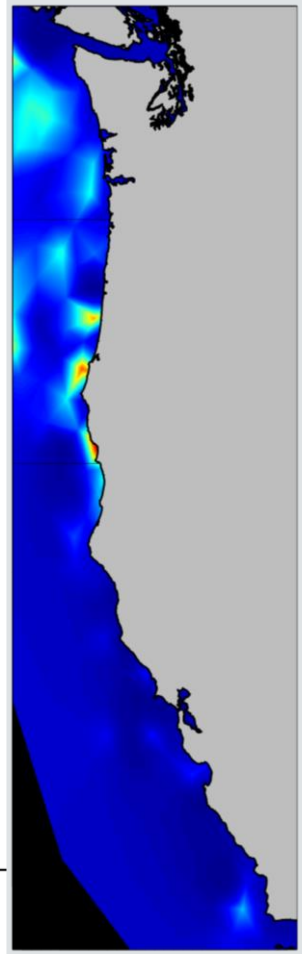


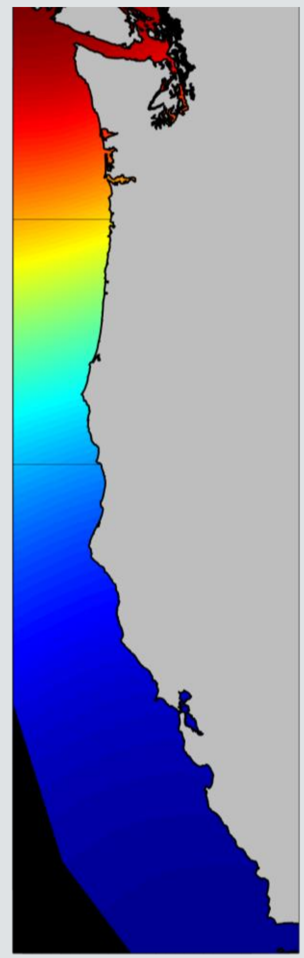
1. Results Binomial

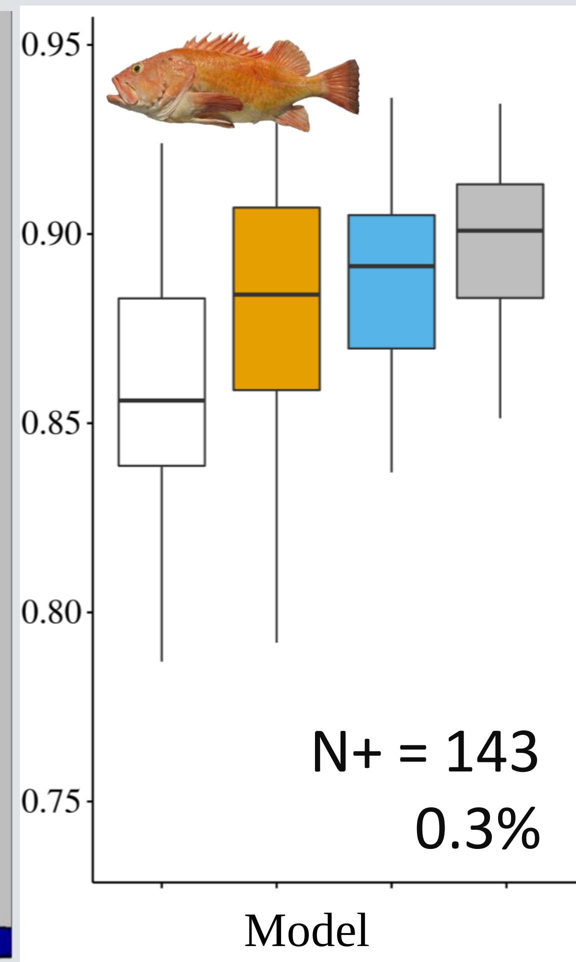
Generally: GLM < GAM < GMRF < RF

Less clear for rarer species



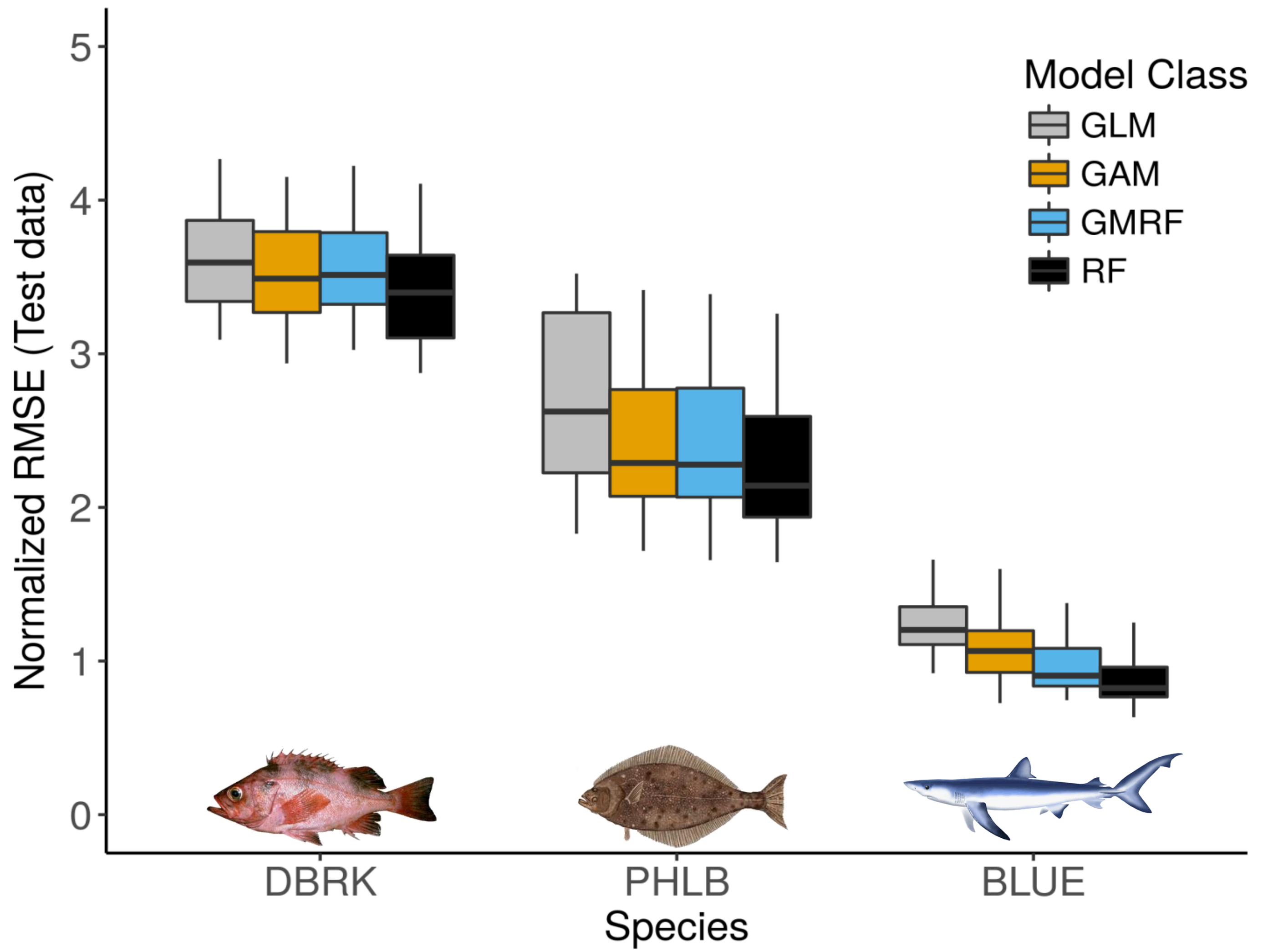






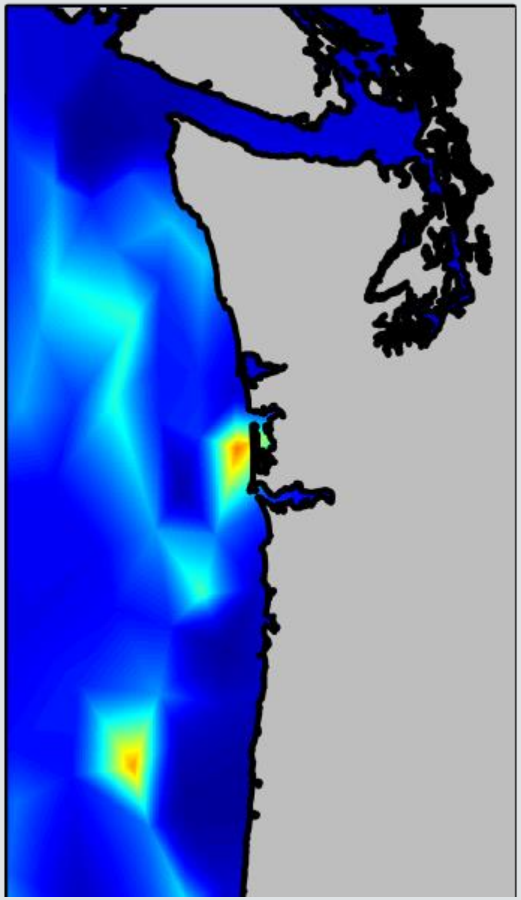
1. Results Binomial

Generally: GLM < GAM < GMRF < RF



1. Results Positive

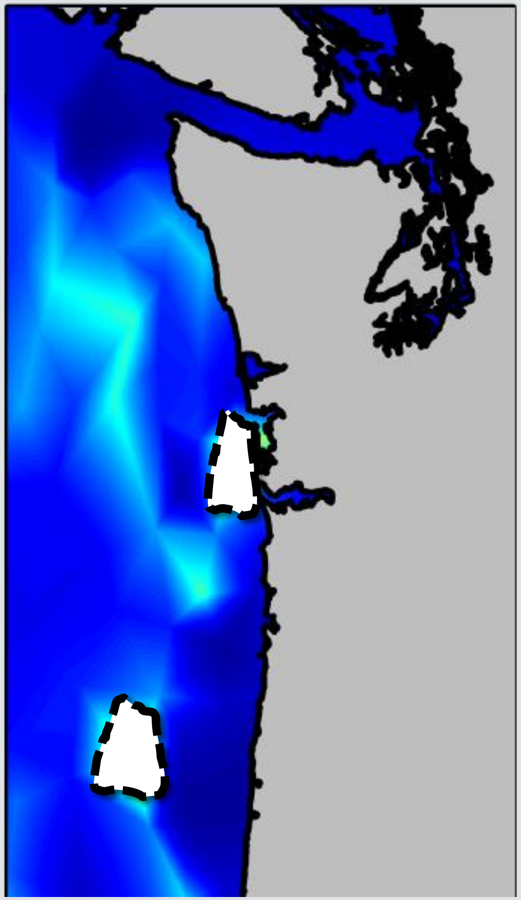
Q: How much bycatch can they prevent?



Crude management simulation:

1. Predict bycatch risk at test locations

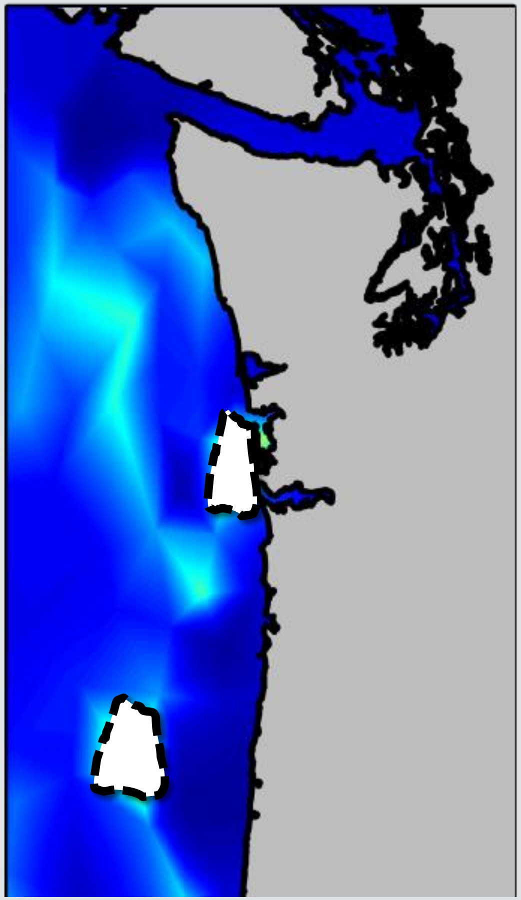
Q: How much bycatch can they prevent?



Crude management simulation:

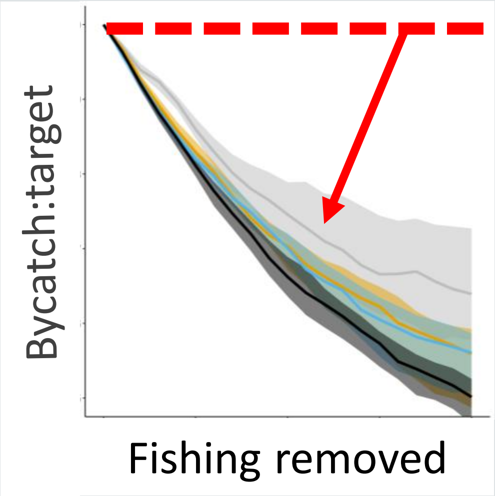
1. Predict bycatch risk at test locations
2. Remove X% of fishing effort with highest bycatch risk

Q: How much bycatch can they prevent?



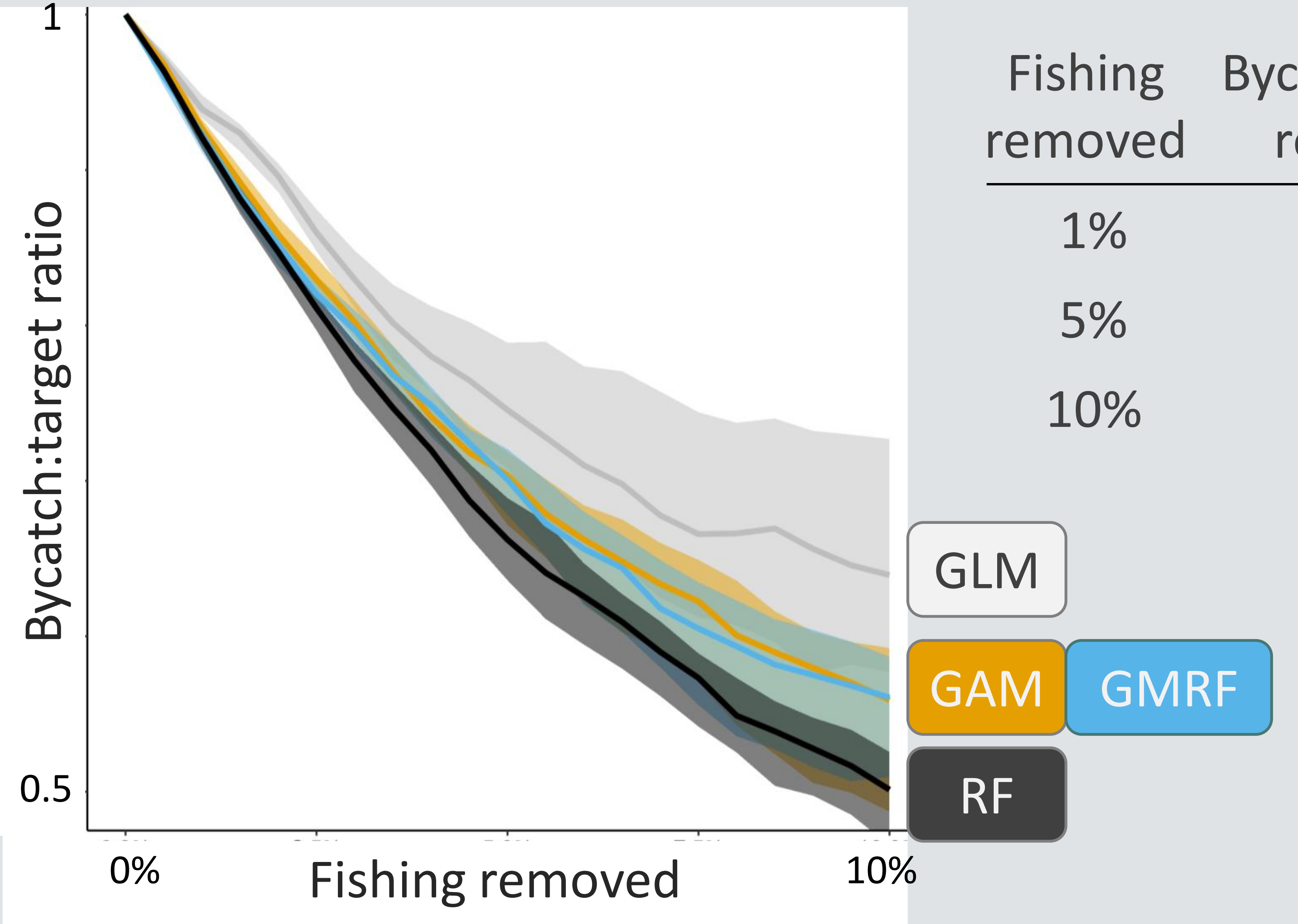
Crude management simulation:

1. Predict bycatch risk at test locations
2. Remove X% of fishing effort with highest bycatch risk
3. Calculate “prevented” bycatch and target catch (bycatch:target ratio)



1. Methods: evaluation

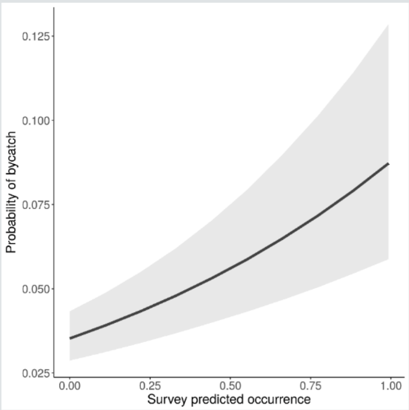
Q: How much bycatch can they prevent?

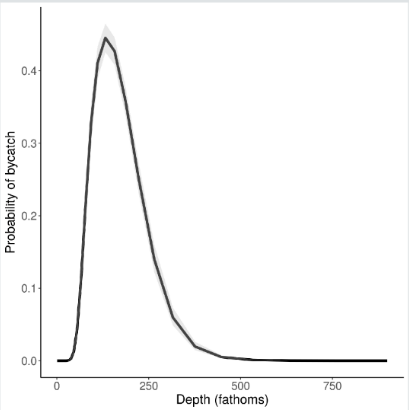


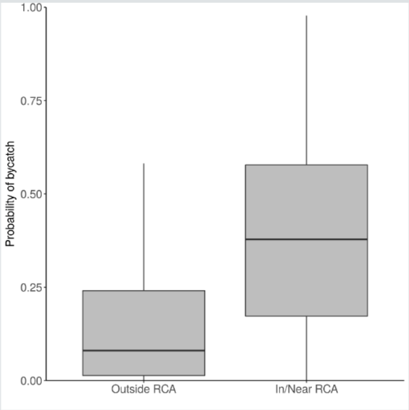
Fishing removed	Bycatch:target reduction
1%	8%
5%	34%
10%	50%

Covariate effects

GMRF







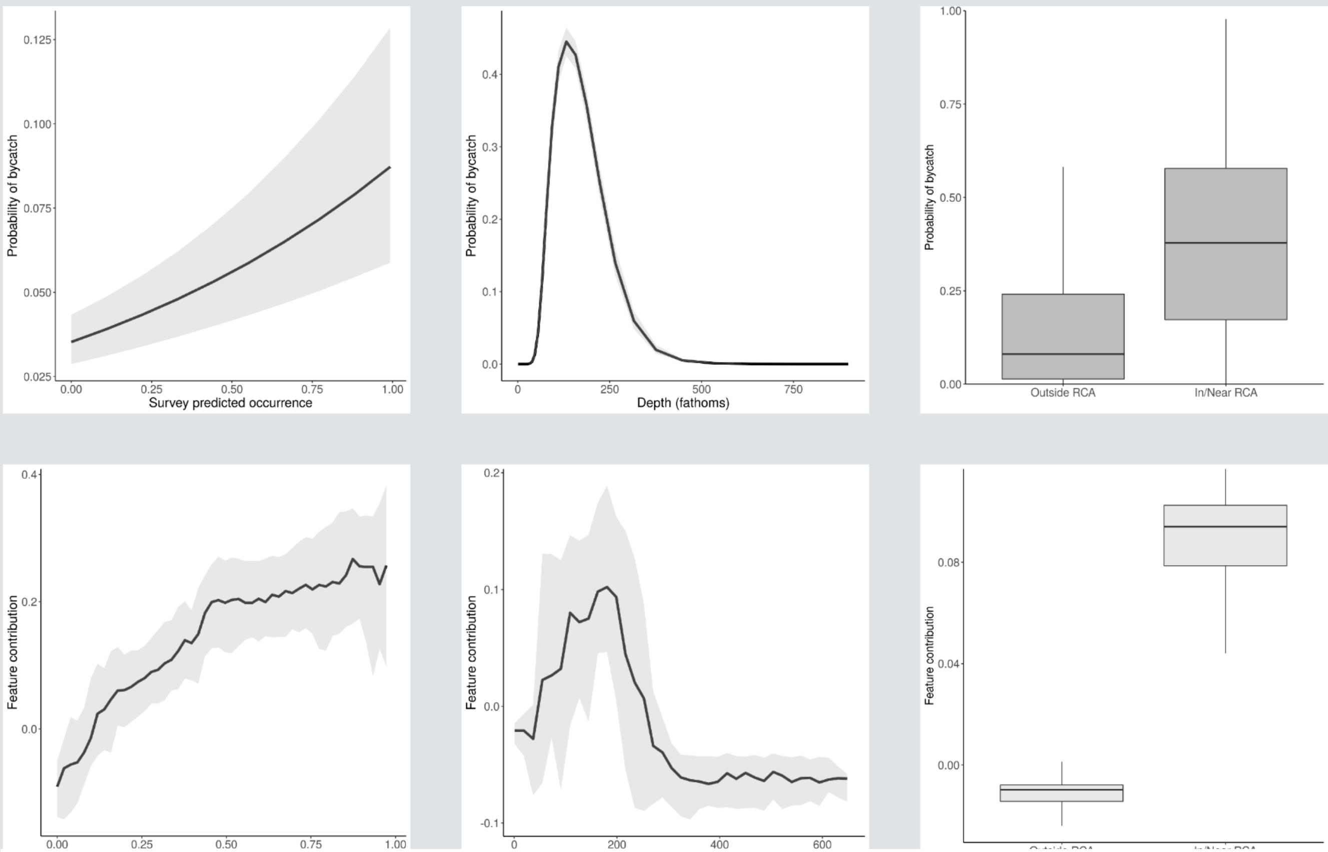
PredOccSurvey | Depth | In/near RCA

Palczewksa (2013), Welling (2016)

Covariate effects

GMRF

RF



PredOccSurvey

Depth

In/near RCA

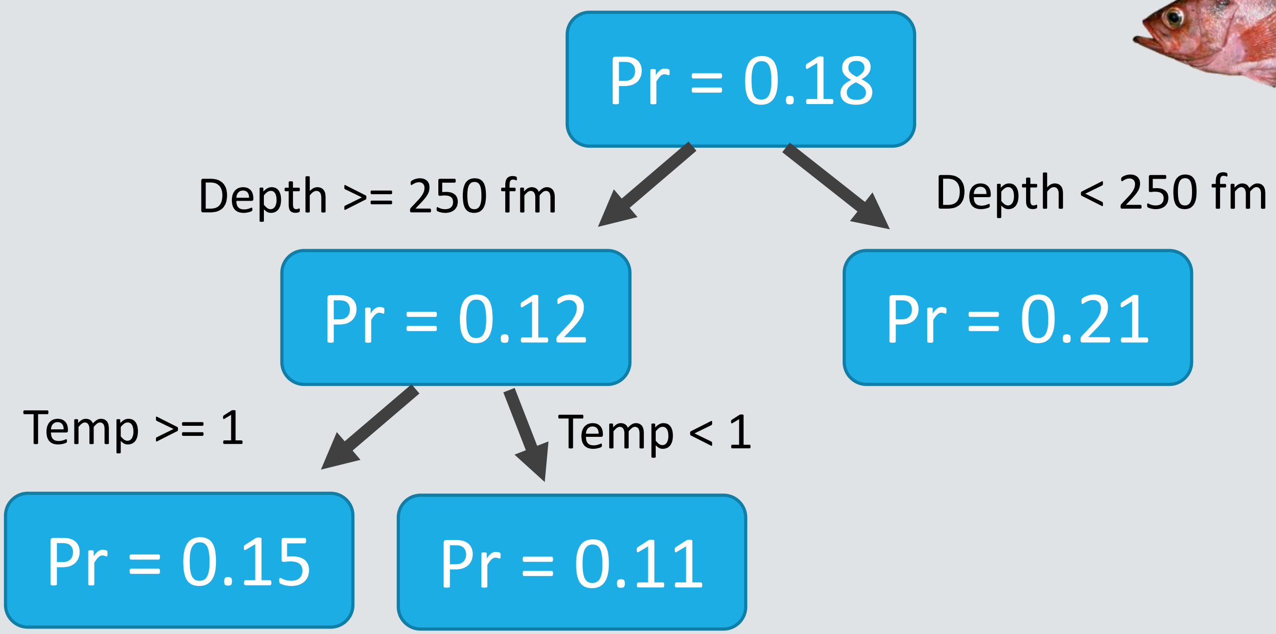
How?

Palczewksa (2013), Welling (2016)

How do random forests work?

Single decision tree:

Low bias, high variance model (overfit)



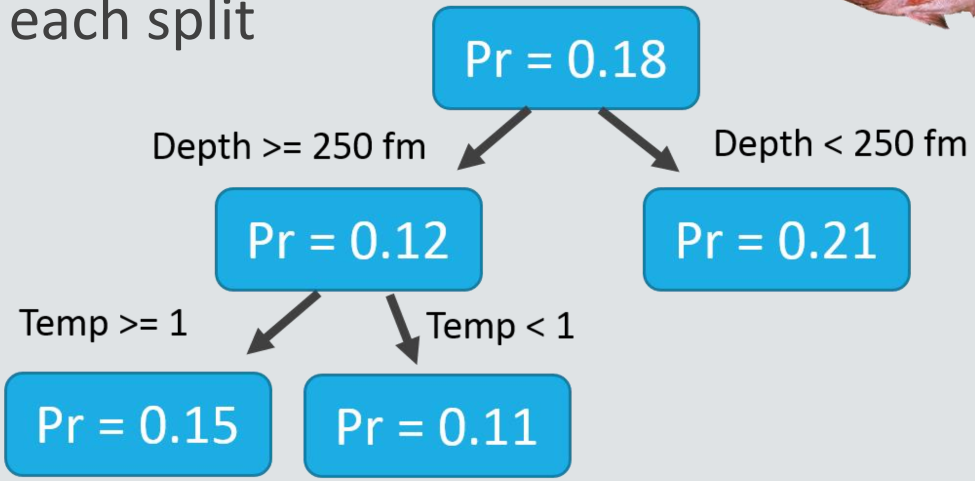
Kuhn & Johnson (2013)

How do random forests work?

Idea: average across many, uncorrelated trees

$$E[MSE] = Model\ Bias^2 + Model\ Variance + noise$$

1. *Bagging:* fit each tree on a **B**ootstrap sample (~63%) of the data, then **Agg**regate across trees (~1000+)
2. Only consider a *random subset* (~P/3) *of covariates* at each split



Kuhn & Johnson (2013)

Covariate effects with RF

What is a “feature contribution”??

Pr = 0.18

Depth >= 250 fm

Depth < 250 fm

Pr = 0.12

Pr = 0.21

Temp >= 1

Temp < 1

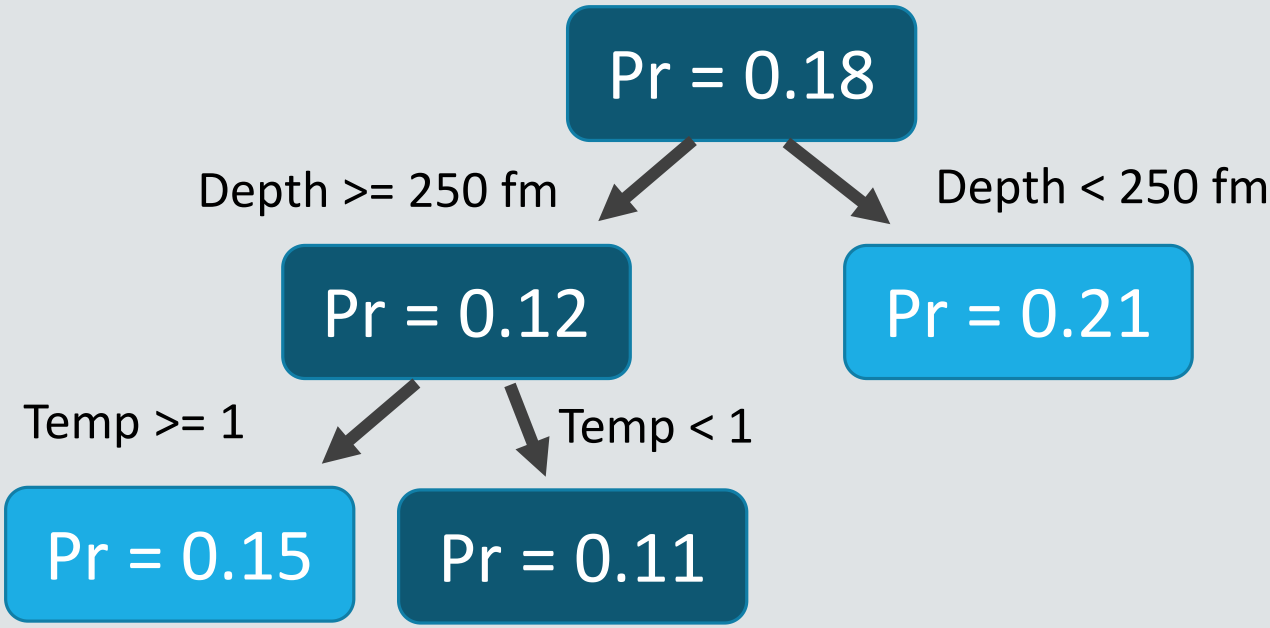
Pr = 0.15

Pr = 0.11

Palczewksa (2013), Welling (2016)

Covariate effects with RF

What is a “feature contribution”??

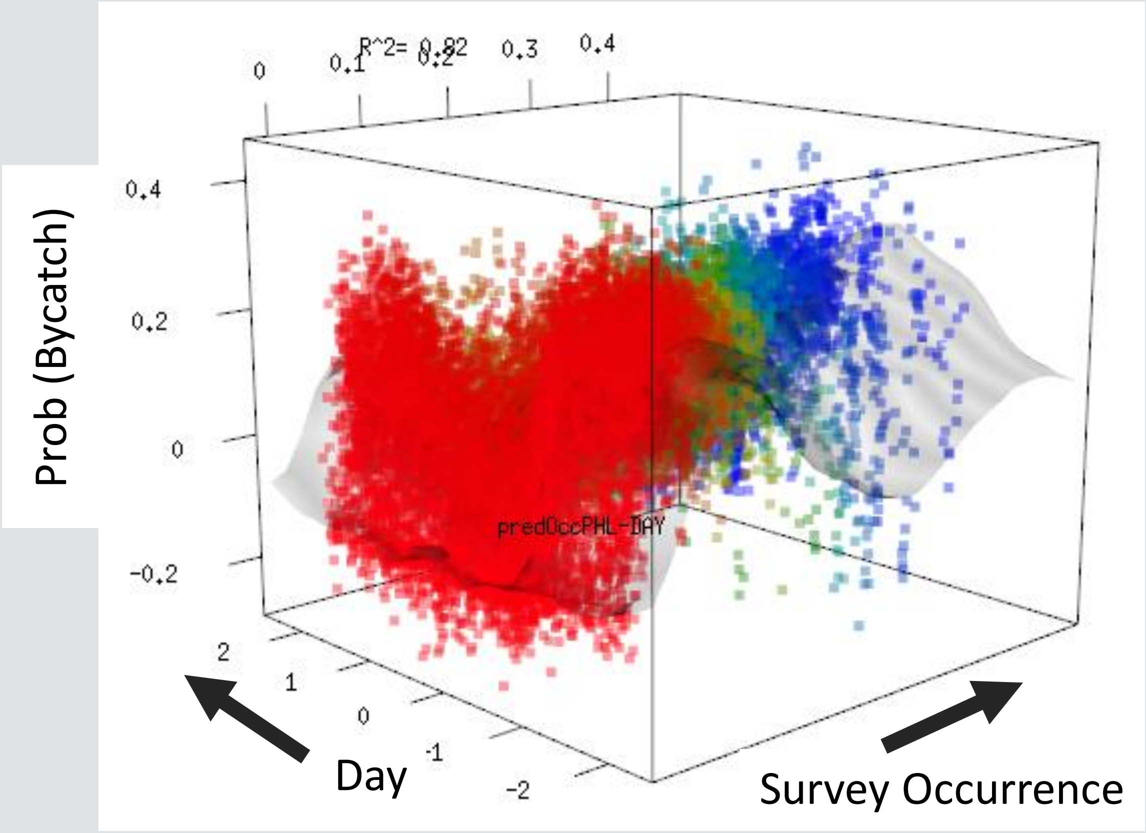


$\text{Prediction}_i = 0.11 = 0.18 - 0.06 \text{ (Depth)} - 0.01 \text{ (Temp)}$

Palczewkska (2013), Welling (2016)

Covariate interactions with RF

Catchability varies by Julian Day



#2: Total bycatch estimates

Need estimates of total bycatch / discards

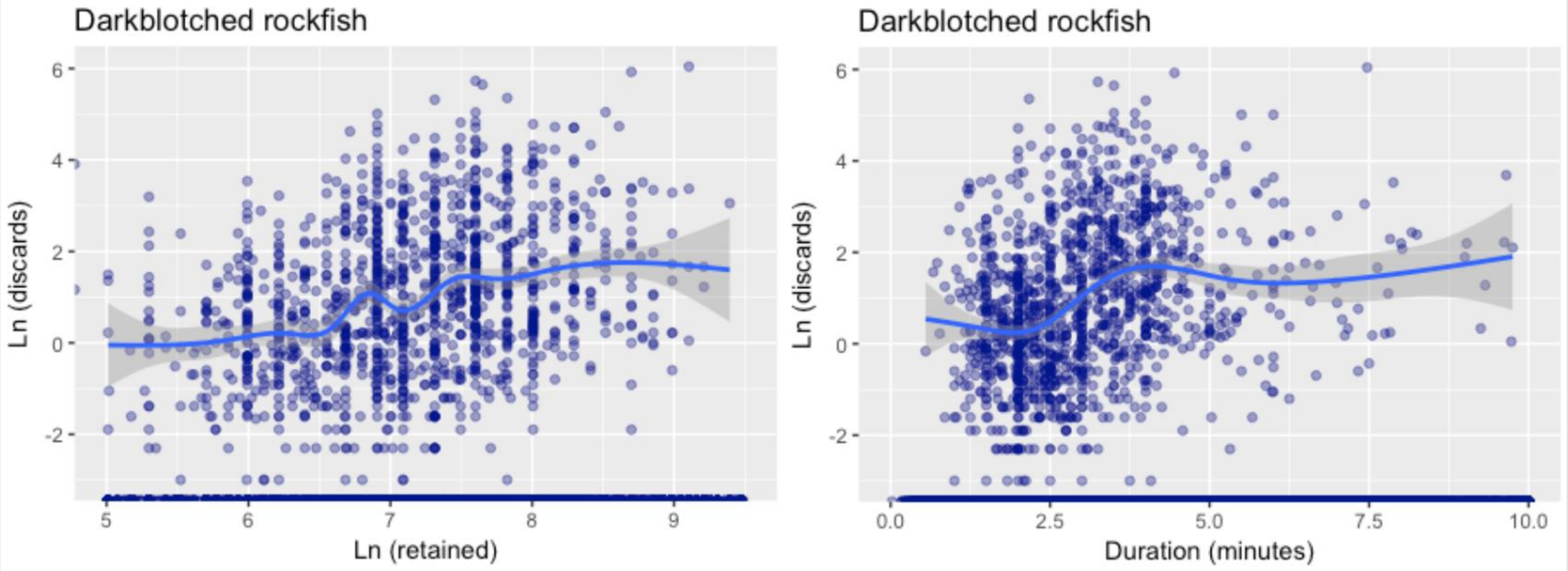
- Rarely observe 100% of fishing
- Often observe ~20%

#2: Total bycatch estimates

"Ratio estimator":

$$B_{unobs} = T_{unobs} \frac{B_{obs}}{T_{obs}}$$

Assumes bycatch prop. to target catch / effort

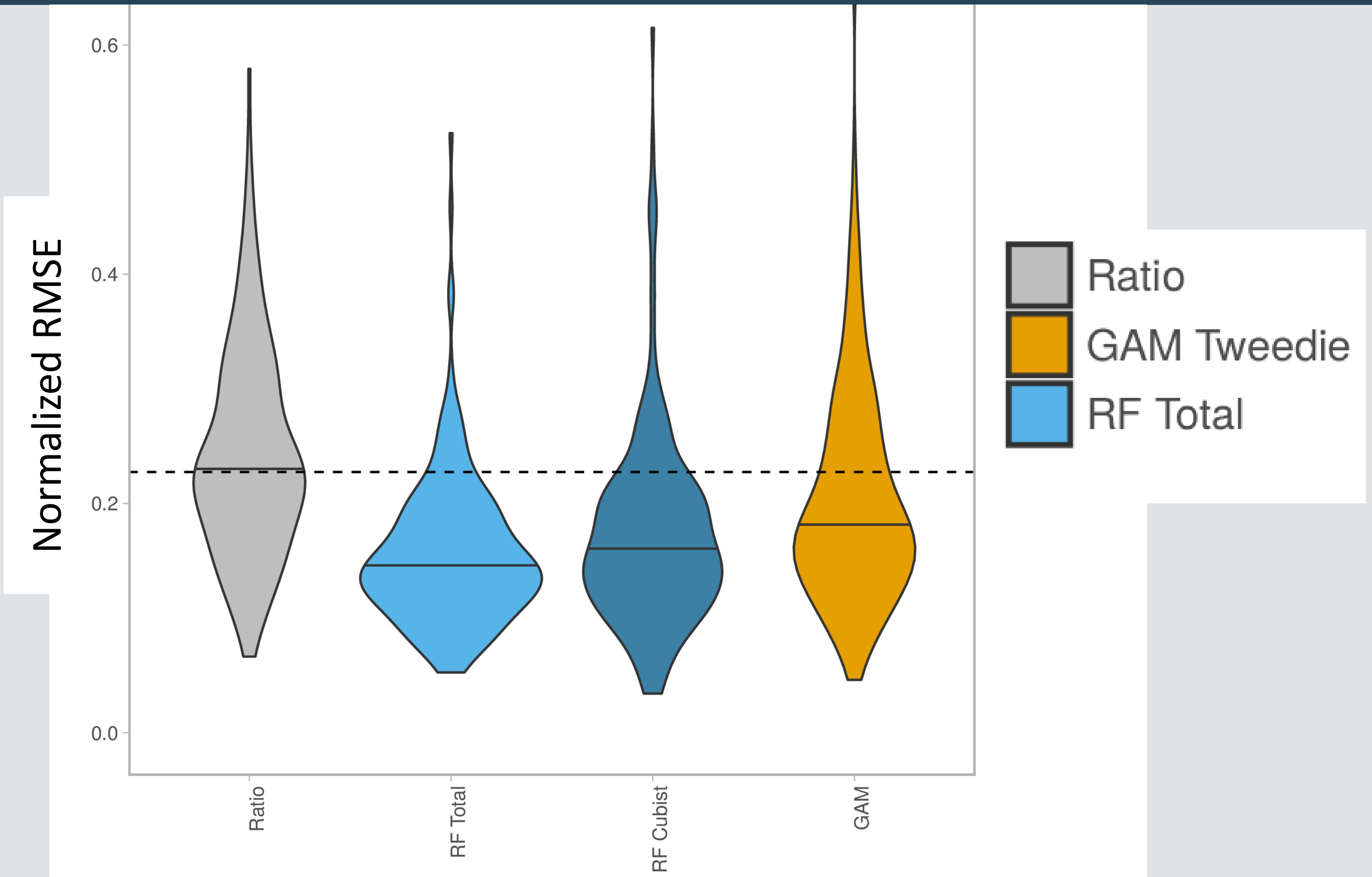


Use a spatial model instead

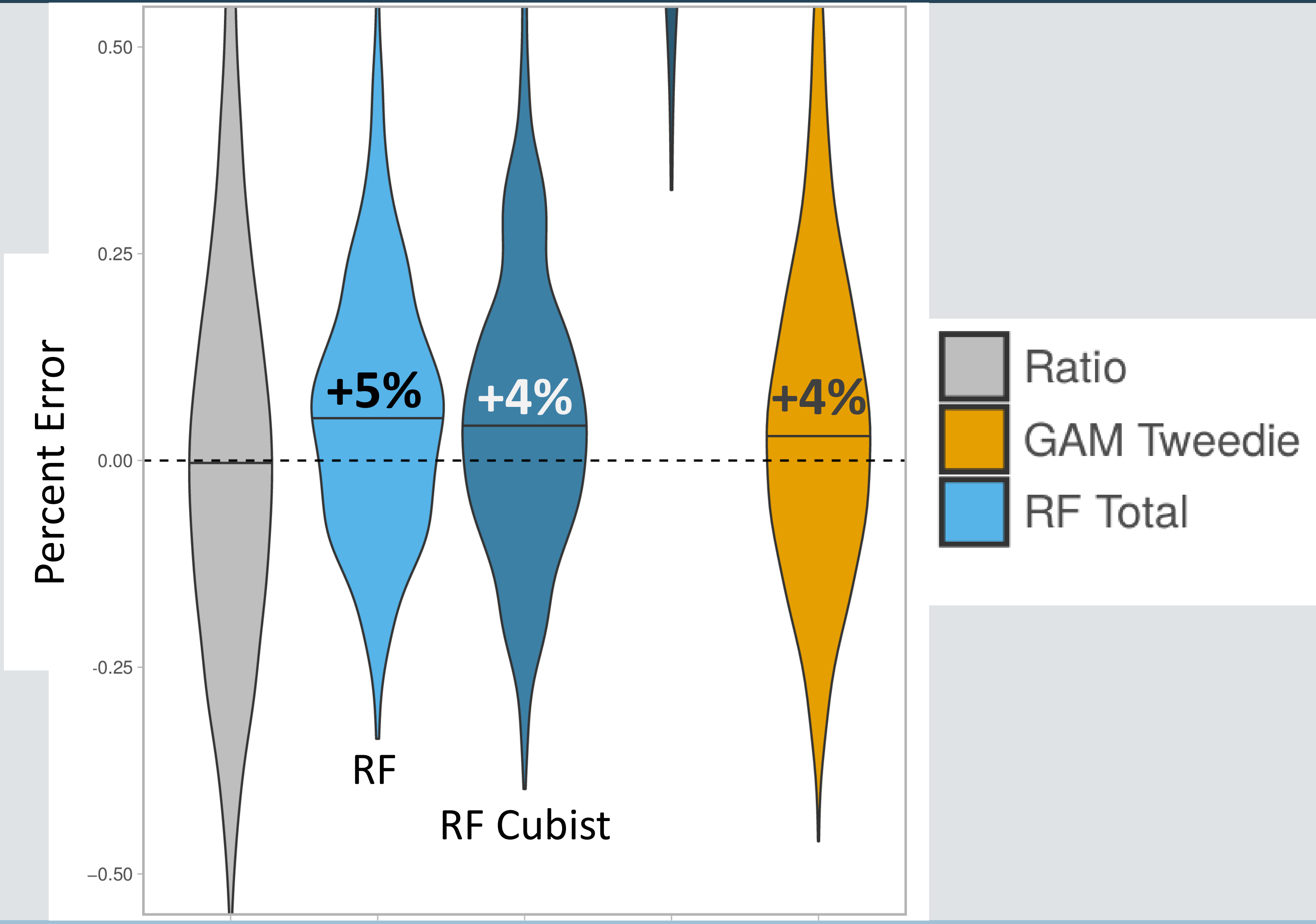
Cross-validation using dataset with 100% coverage:

1. Choose 20% observed trips
2. Fit spatial model
3. Predict at 80% unobserved
4. Compare sum(predictions) to ratio estimator

Spatial models = lower error

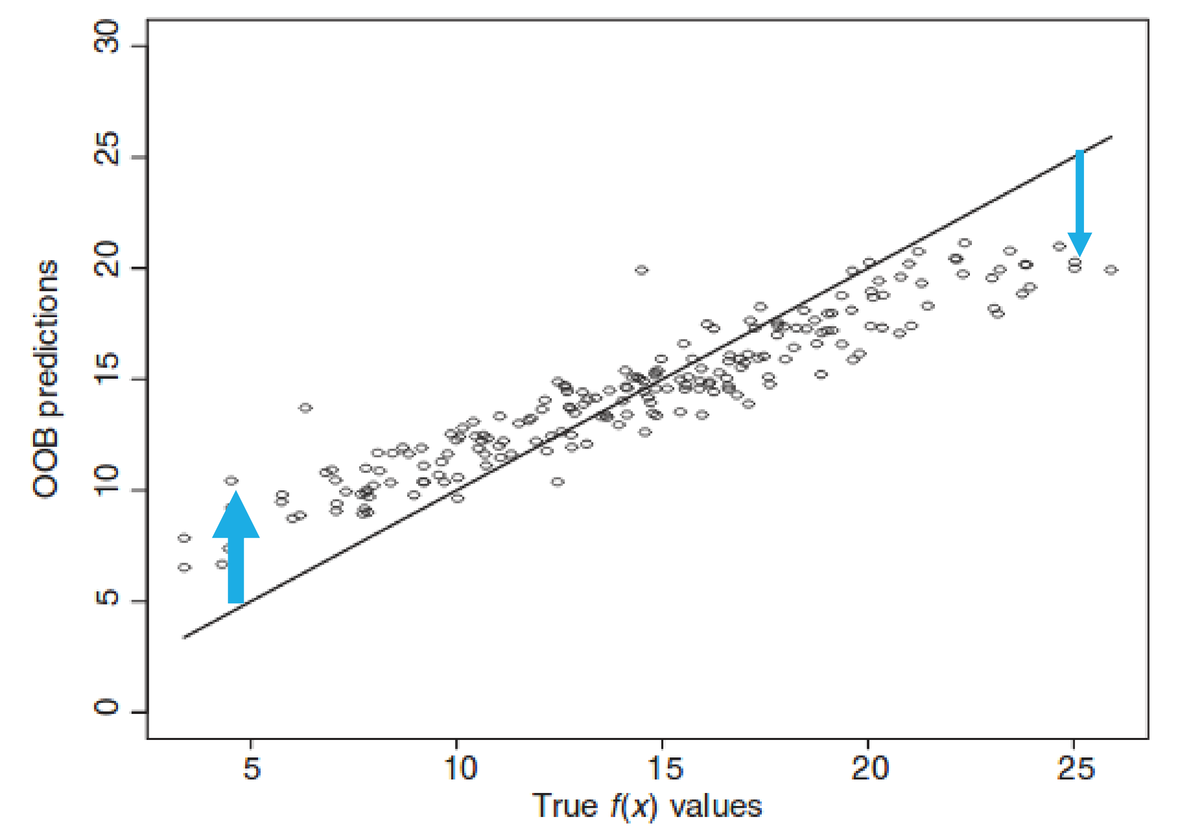


… bias in spatial model estimates



Why are random forests biased?

1. Extreme values modeled using average of less-extreme points → Regression to the mean



2. Bycatch distribution is right-skewed

Zhang & Lu (2012)

Thoughts on RF bias

Bias correction methods:

- Fit linear model in nodes instead of mean (‘Cubist’) --
- Fit second model on RF residuals (Xu 2013)

Bycatch estimates (*absolute* abundance) vs.

CPUE standardization (*relative* abundance)

#3: CPUE data

Eastern Pacific Ocean yellowfin tuna

- 2000-2009 catch + effort
- 1-deg lat/lon bins

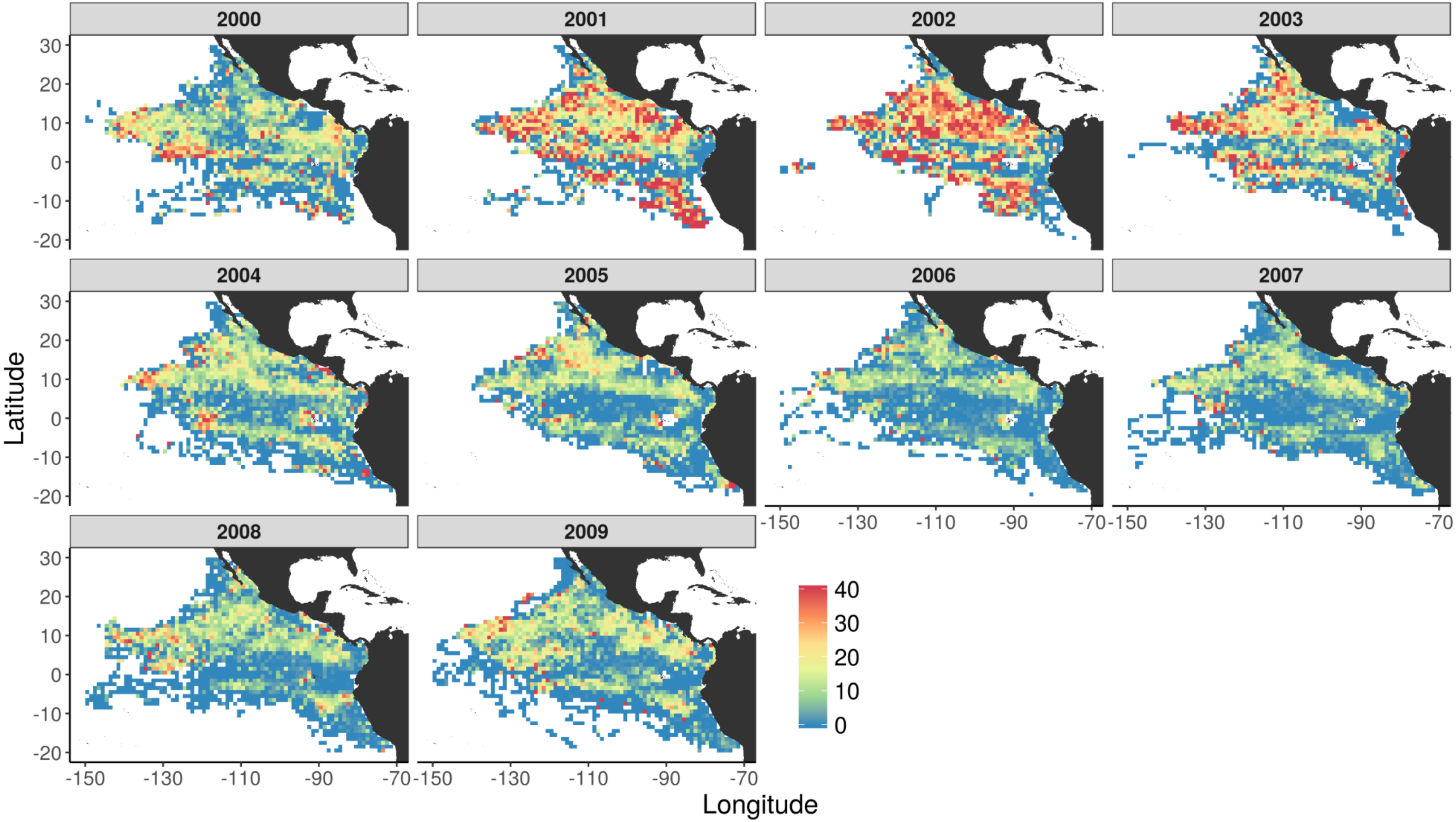
Model:

- 2000-2009 catch + effort
- 1-deg lat/lon bins

‘ranger’ `ranger(cpue ~ lat + lon + year, …)`

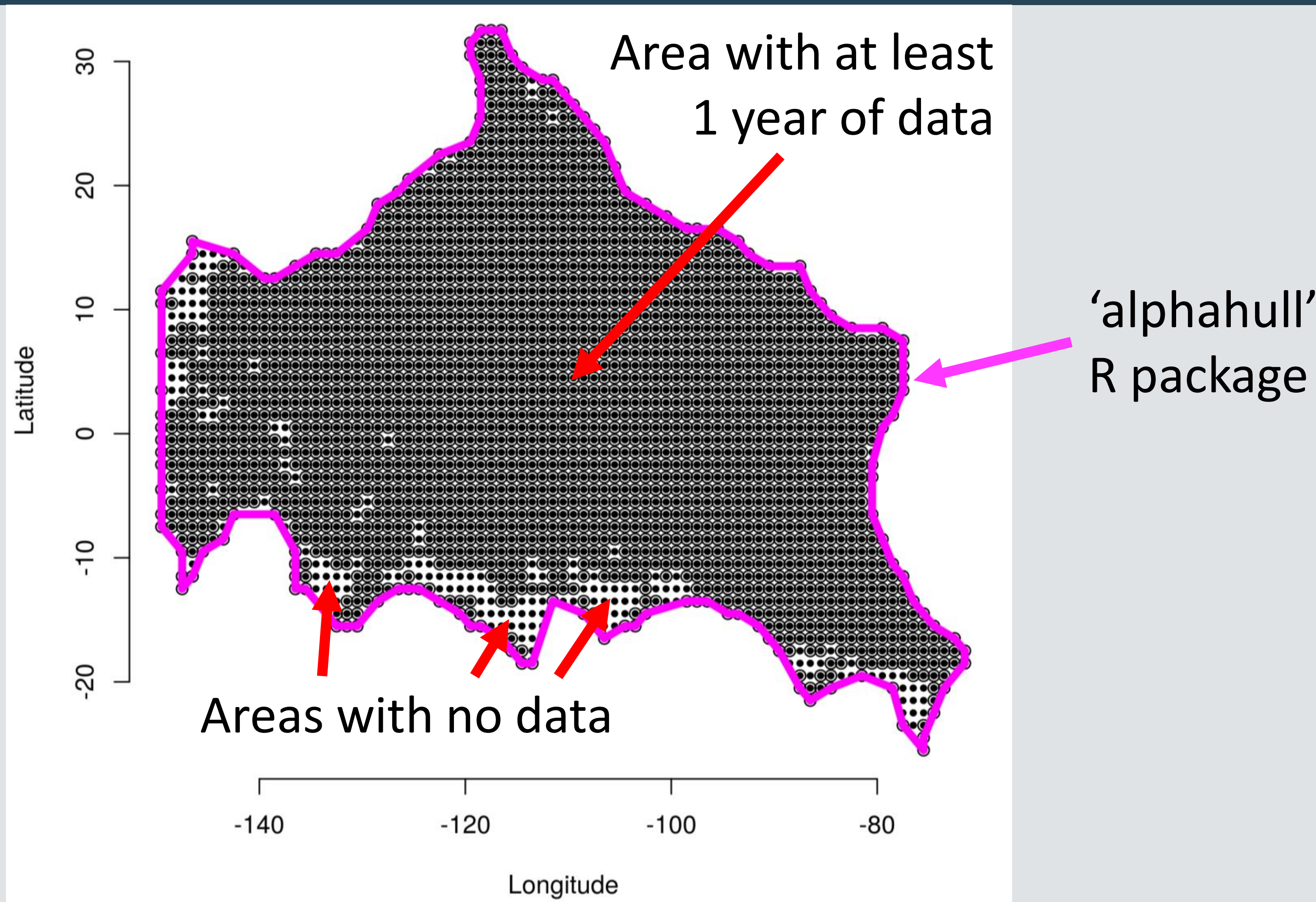
‘grf’ `regression_forest(dat[,covar], Y=dat$cpue, …)`

CPUE data

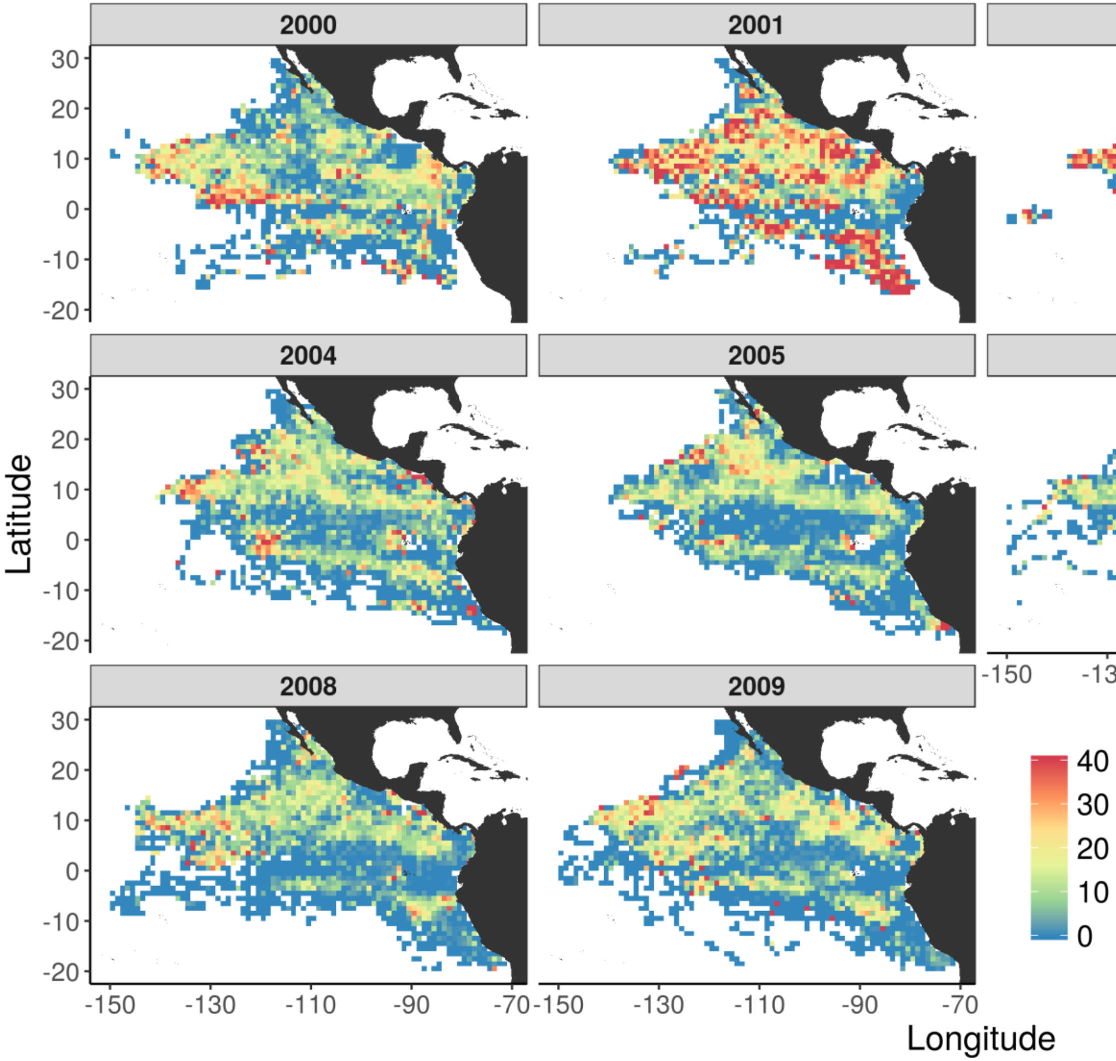


2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
Latitude
Longitude
30
20
10
0
-10
-20
-150
-130
-110
-90
-70
40
30
20
10
0

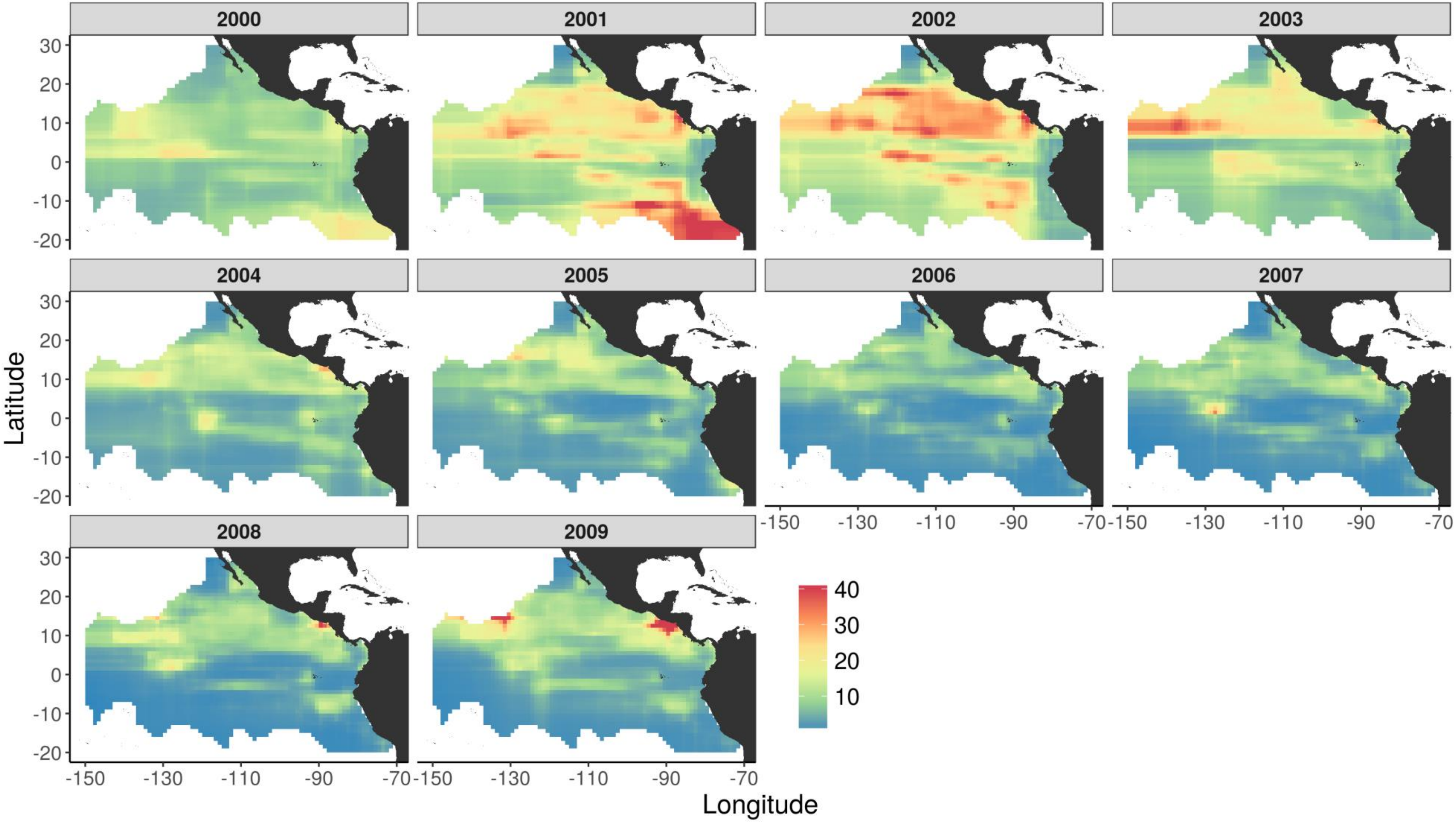
Create prediction grid



CPUE data

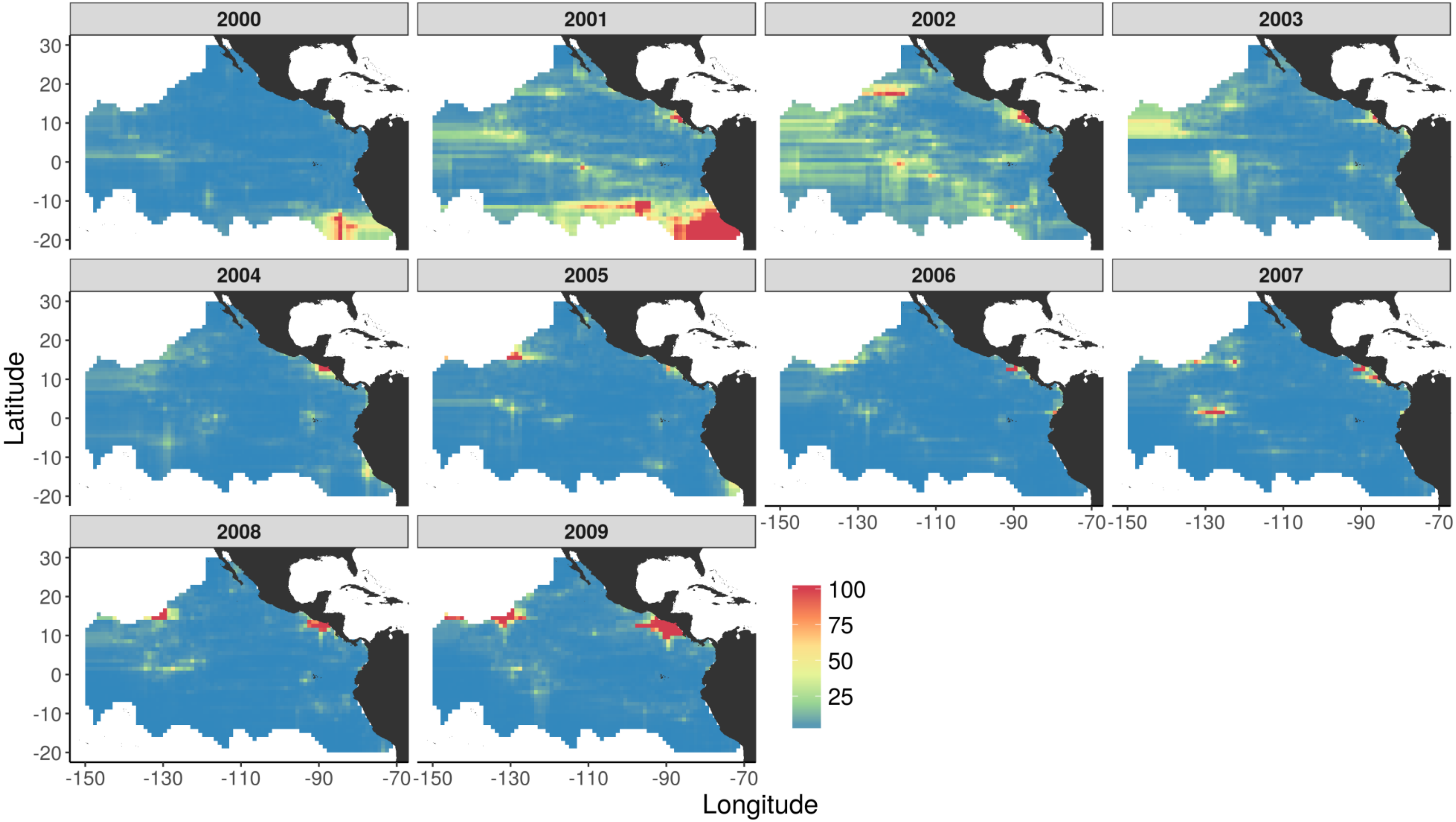

2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
30
20
10
0
-10
-20
Latitude
-150
-130
-110
-90
-70
Longitude
40
30
20
10
0

Predicted mean(CPUE)



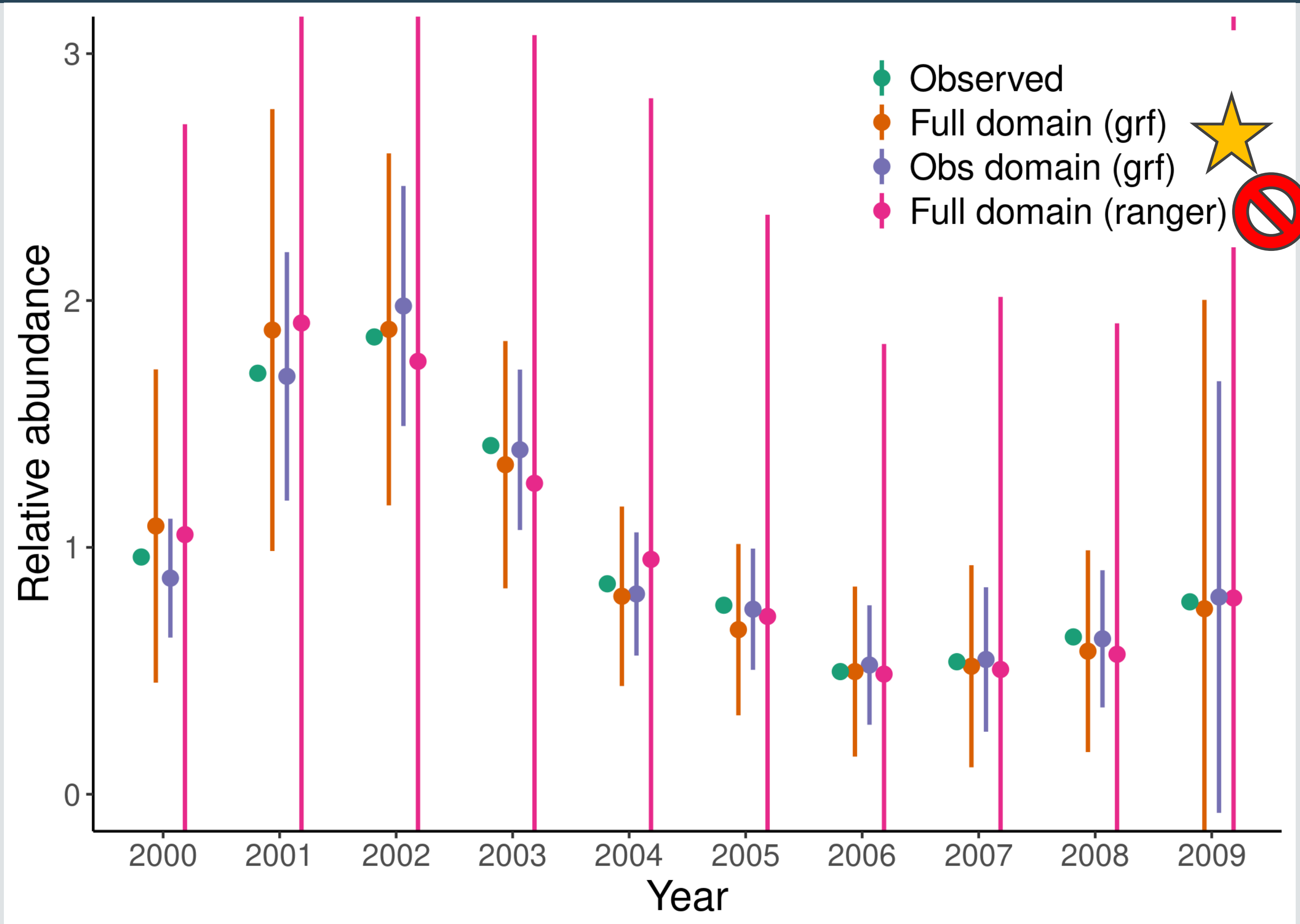
3. Results

Predicted Var(CPUE)



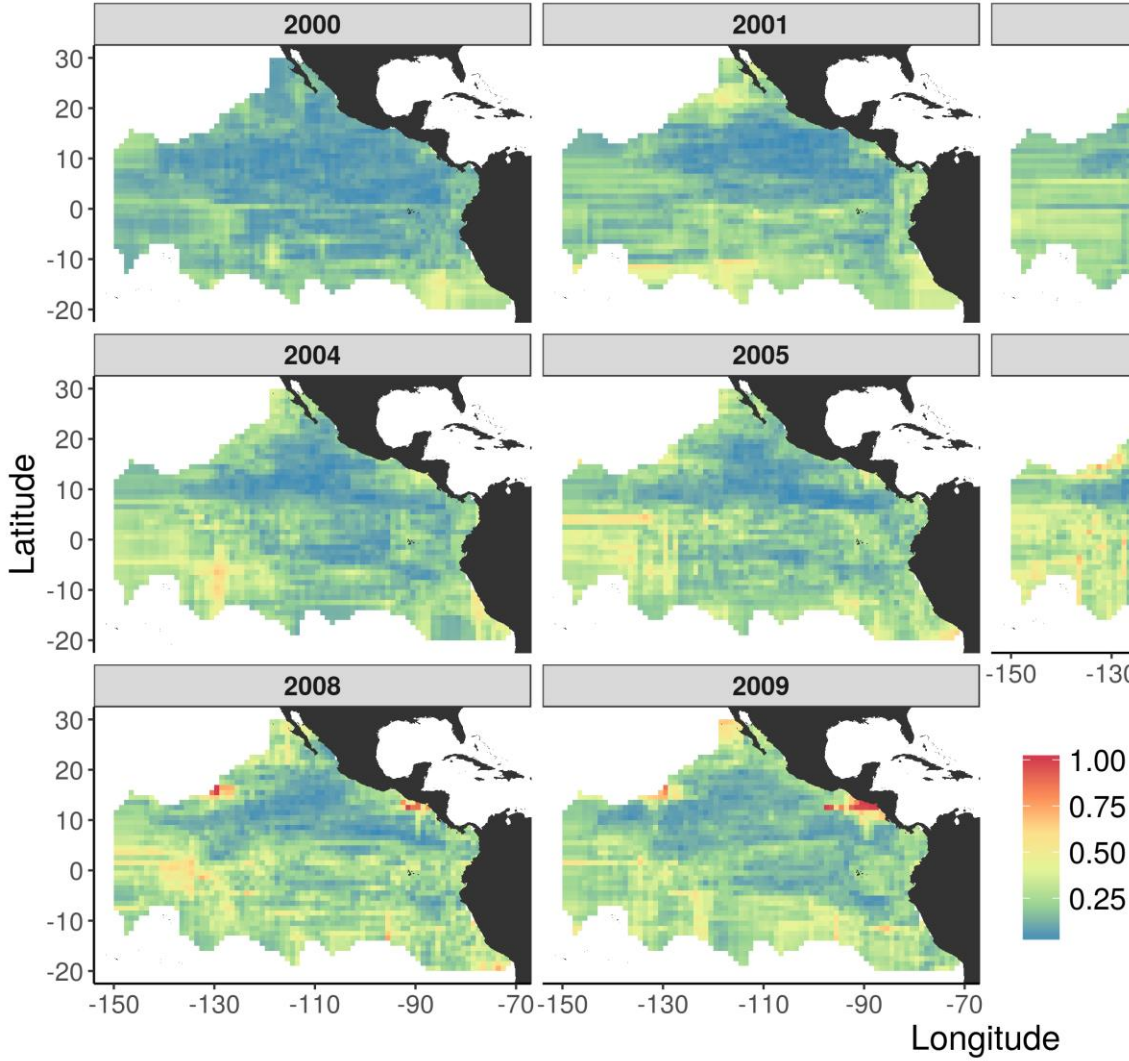
3. Results

Relative abundance trend



3. Results

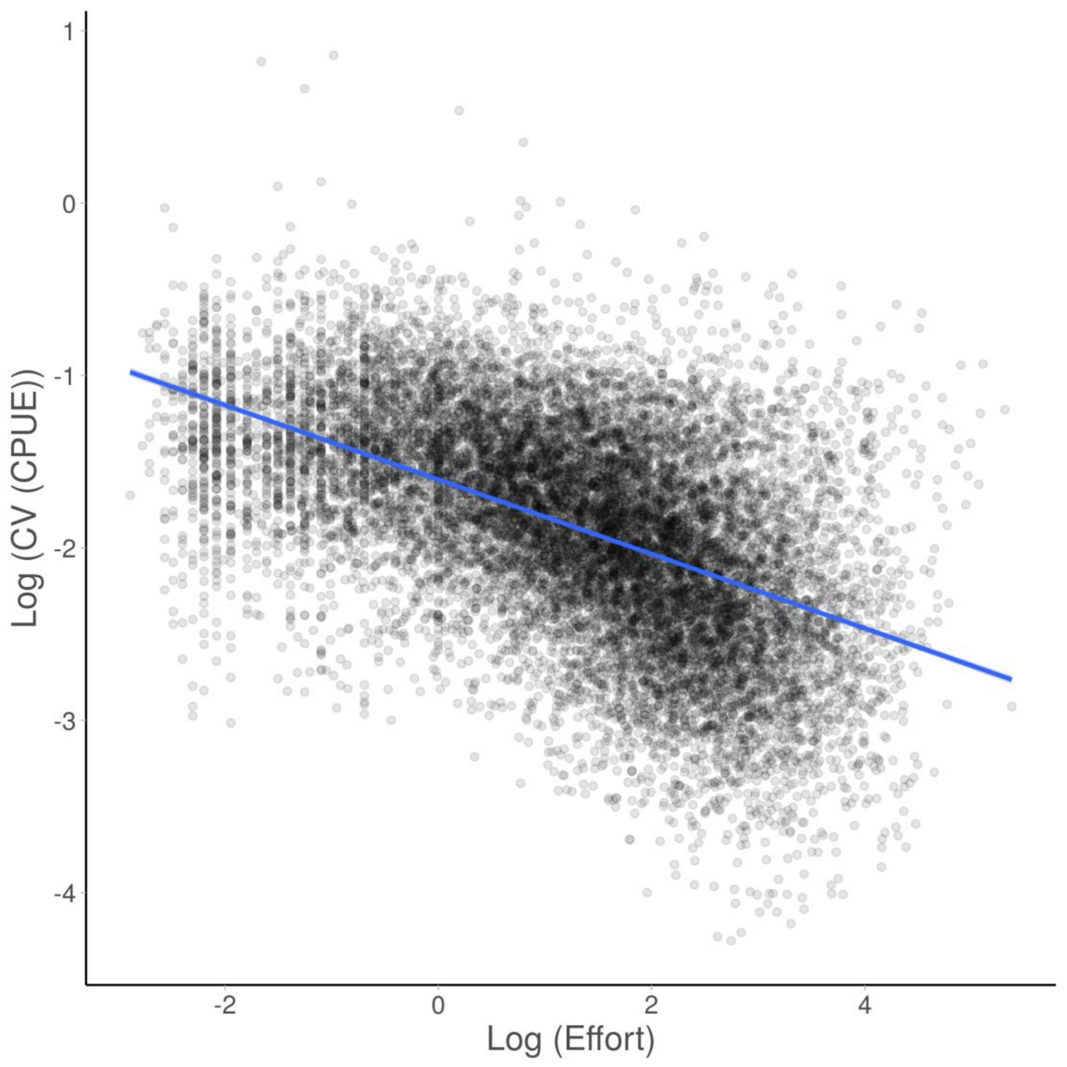
Predicted CV(CPUE)



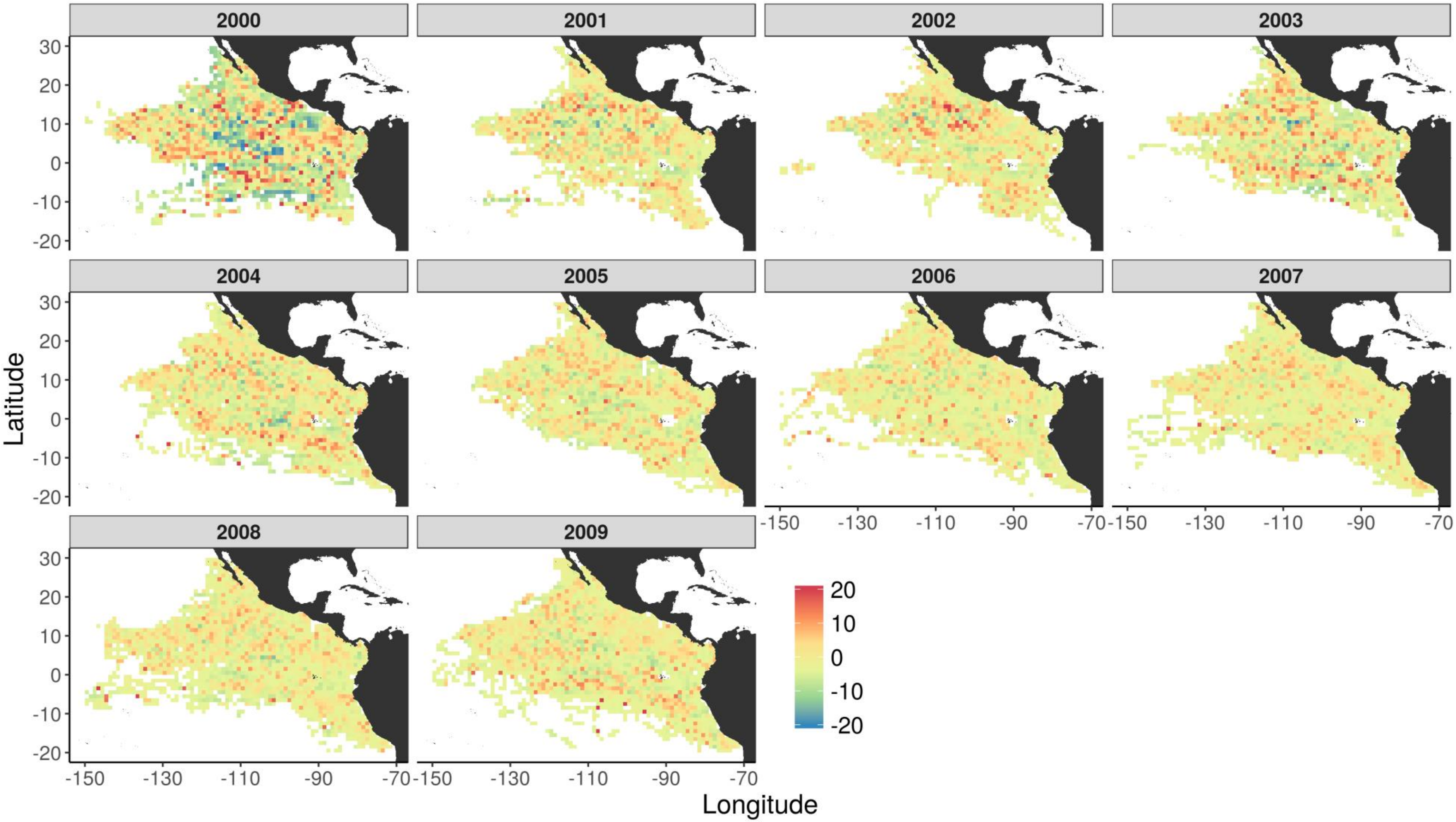
Reasonable scale ⭐

Reasonable pattern ⭐

log(CV) vs. log(Effort)

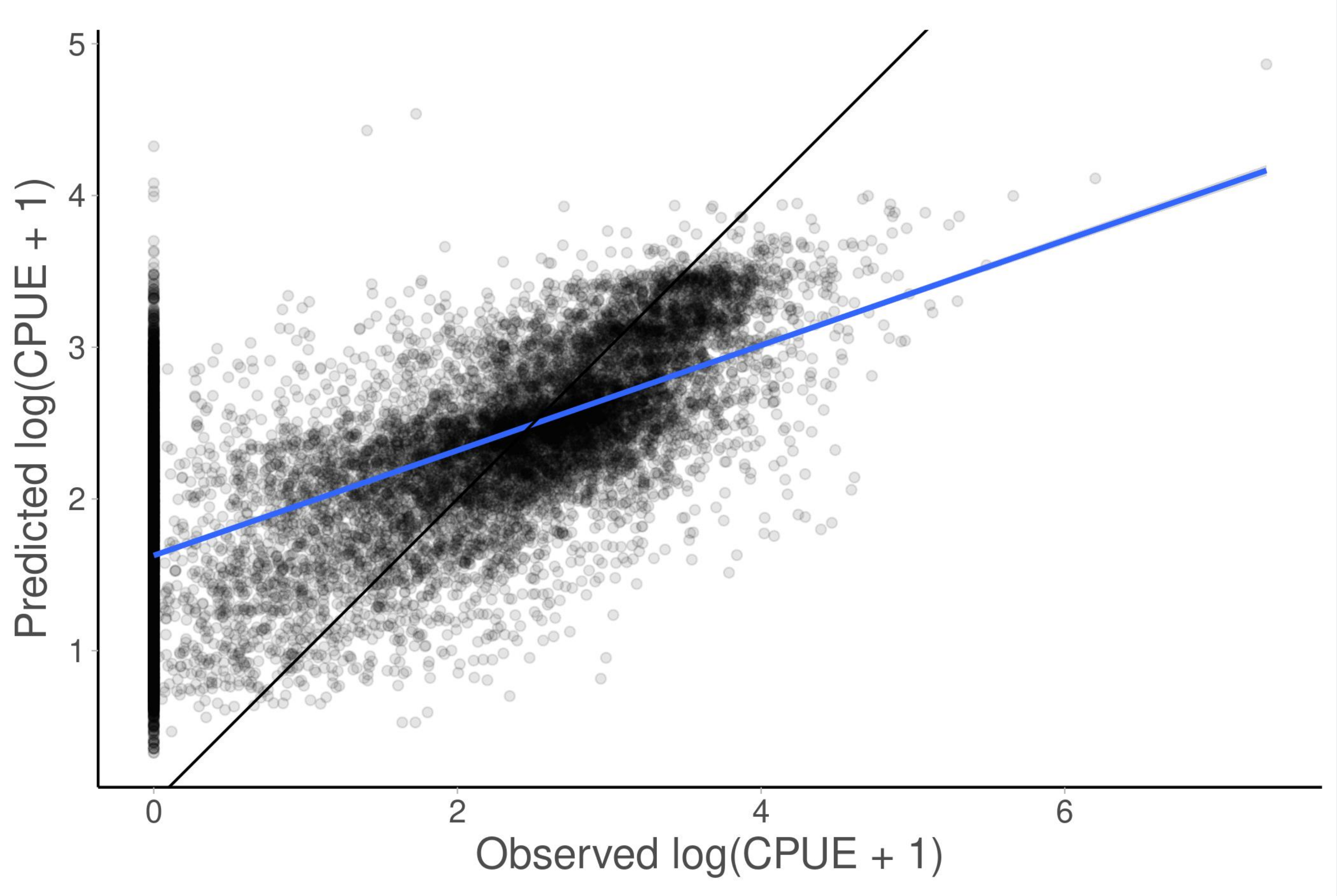


Standardized residuals



3. Diagnostics

Bias (regression to the mean)



Uncertainty estimates

Need *covariance* between spatiotemporal predictions

Rapidly evolving… 34,336 citations Breiman (2001)

1. Quantile regression forests – prediction quantiles
   (‘ranger’, ‘grf’, Meinshausen 2006)
2. Jackknife & infinitesimal jackknife – standard error
   (‘ranger’, Wager et al. 2014)
3. U-statistics – asymptotically normal variance estimate
   (‘surfin’, Mentch & Hooker 2016)
4. Generalized random forests – asymp. normal variance est.
   (‘grf’, Athey et al. 2017)
5. Bayesian additive regression trees – full posterior inference
   (‘bayesMachine’, ‘dbarts’, ‘BART’, Chipman et al. 2010)

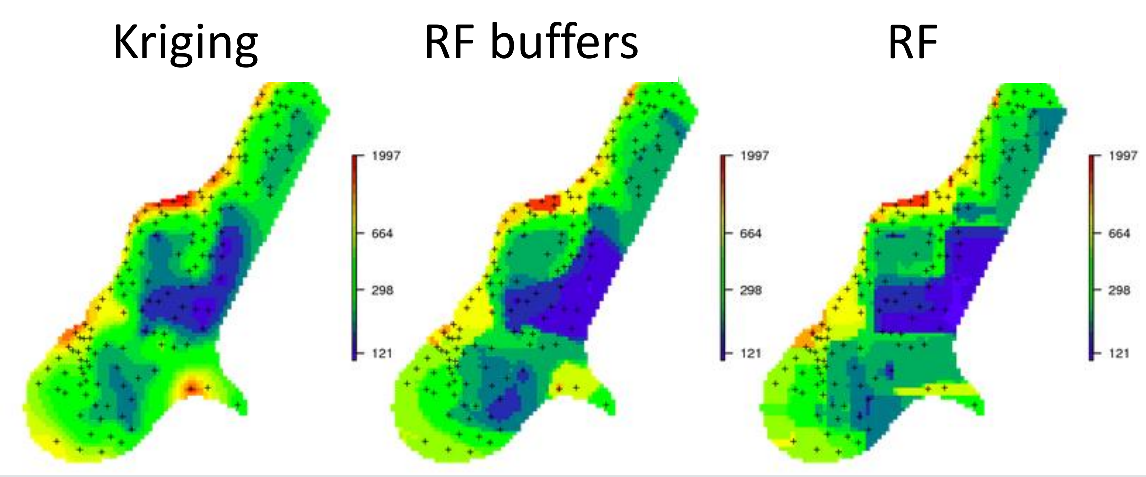
Other thoughts

Multivariate response:

- Include model.matrix as covariates:

```
levels(Data_Geostat$spp) <- c("A_stomias", "G_chalcogrammus","H_elassodon")
sp.mat <- data.frame(model.matrix(~ spp - 1, Data_Geostat))
mv.dat <- cbind(Data_Geostat, sp.mat)
rfmv = ranger(Catch_KG ~ Lat + Lon + Year + sppA_stomias + sppG_chalcogrammus + sppH_elassodon,
              data=mv.dat, num.trees=1000, mtry=2, keep.inbag=T, write.forest=T)
```

Buffer distances to smooth predictions:



What we want (from Rick Methot)

- ★ Fast (coding vs. runtime vs. interpretation)
- ★ Replicable (method well-defined, get same answer)
- ★ Robust (insensitive to distributional assumptions, outliers)
- ★ Predictive ability (minimal errors, fill in space/time gaps)
- ★ Covariate effects (nonlinear, interactions)
- -- Uncertainty estimates (with known properties)
- 🚫 Specifiable structure (e.g. correlation through time, biology)
- 🚫 Unbiased (relative vs. absolute abundance)

Discussion

Thank you!

SIO

- Brice Semmens

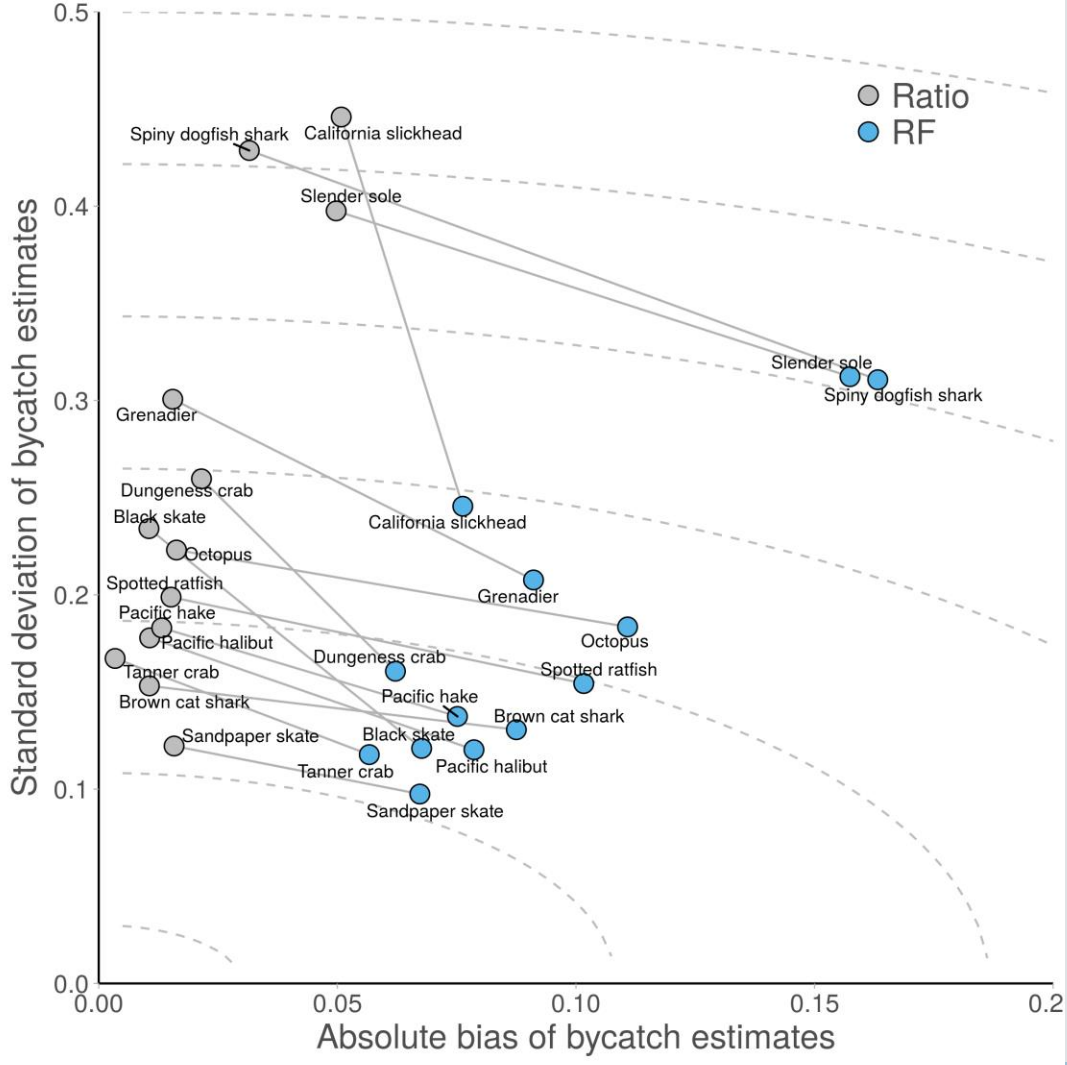
SWFSC

- Tomo Eguchi

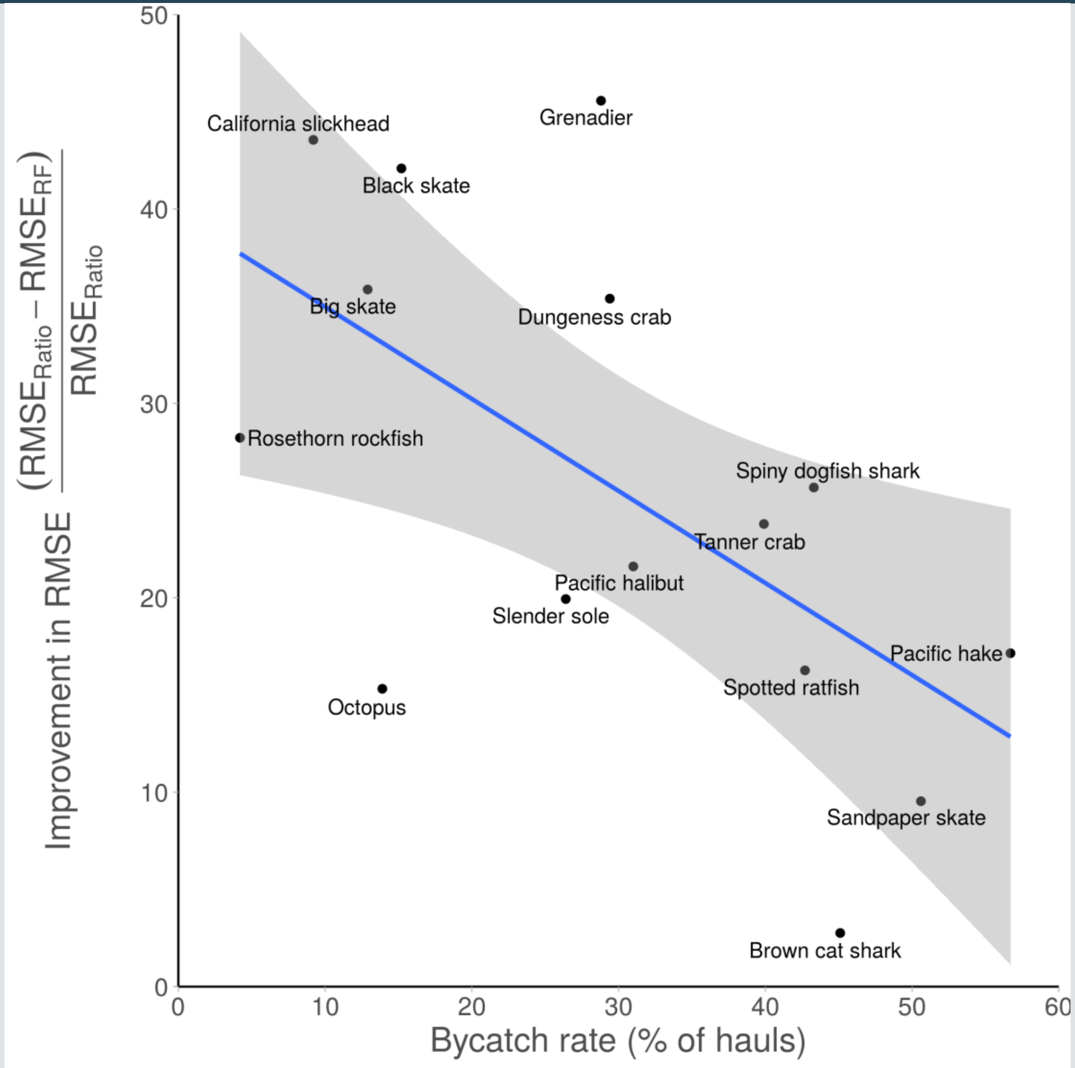
NWFSC

- Eric Ward
- Jim Thorson
- Essential Fish Habitat (Blake Feist)
- West Coast Groundfish Observer Program (Jason Jannot)

Bias-variance tradeoff by species…



More worthwhile for rarer species



2. Results

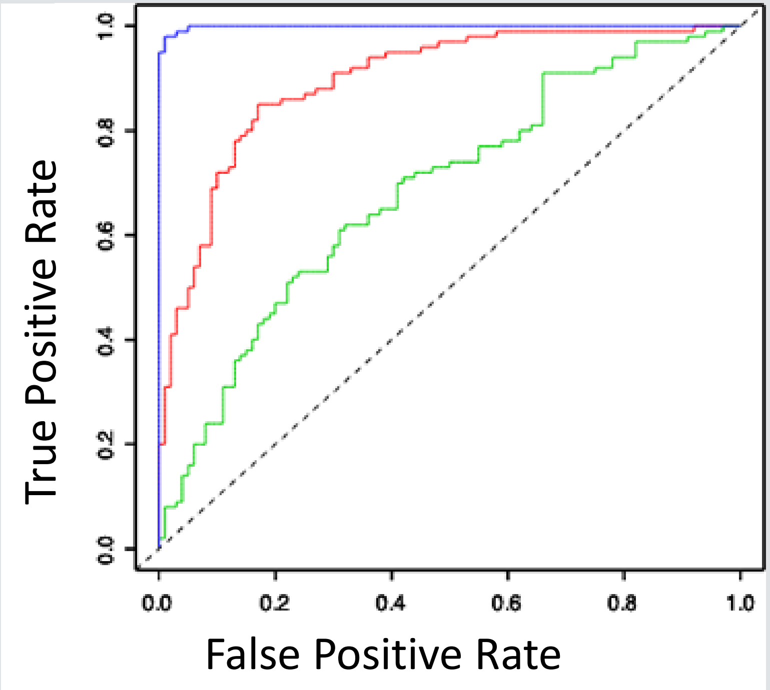
Q1: Which spatial model is best?

Goal: ***prediction***

5-fold cross validation repeated 10x

Binomial

ROC curve (AUC)



ROC curve		AUC
- - -	Worthless	0.5
green	Ok	0.7
red	Good	0.8
blue	Awesome	0.9+

Q1: Which spatial model is best?

Goal: ***prediction***

5-fold cross validation repeated 10x

Binomial

AUC

Positive

RMSE, R^2 (pred – obs)

$$\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

West Coast Groundfish covariates

Binomial

Positive

$\sim$ sst + sst^2 +

depth + depth^2 +

distance to rocky substrate +

size of rocky patch +

in Rockfish Conservation Area +

predicted occurrence (survey) +

day of year +

spatial field

Shelton et al. (2014)

Hawaii Longline covariates

Binomial

Positive

$\sim$ sst + sst^2 +

day of year +

spatial field

Shelton et al. (2014)

RF

+ Better at prediction

+ More complex covariate relationships (incl. interactions)

+ Easier to set up and run

+ Not just a “black box”?

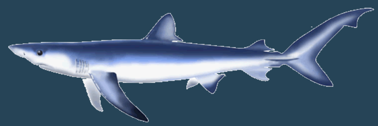
GMRF

+ Statistical inference, marginal posteriors for covariate effects

+ Ability to include observation error

Discussion

Variance of predictions

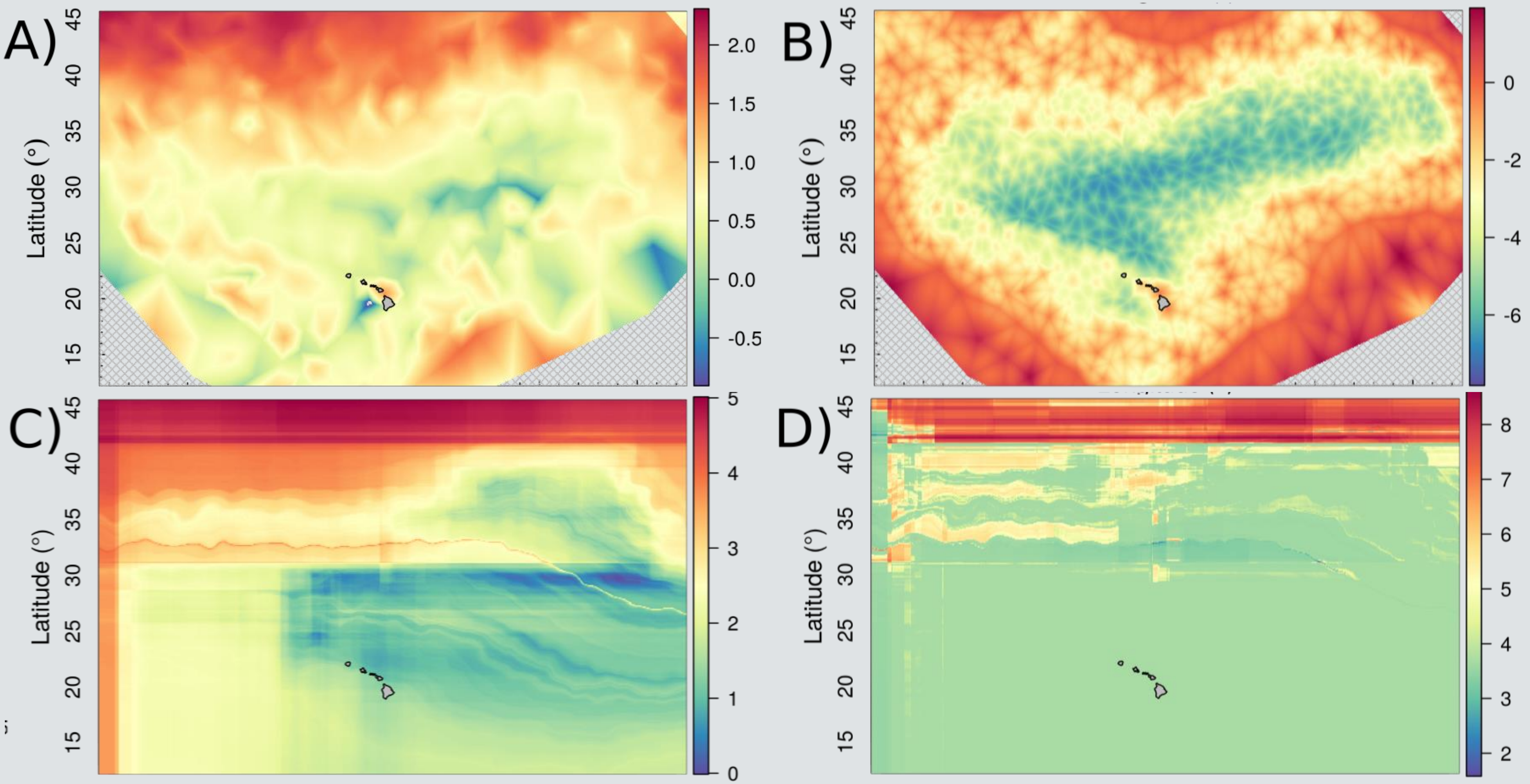


Mean

Variance

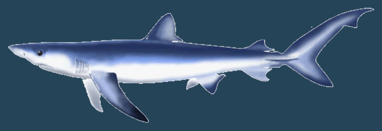
GMRF

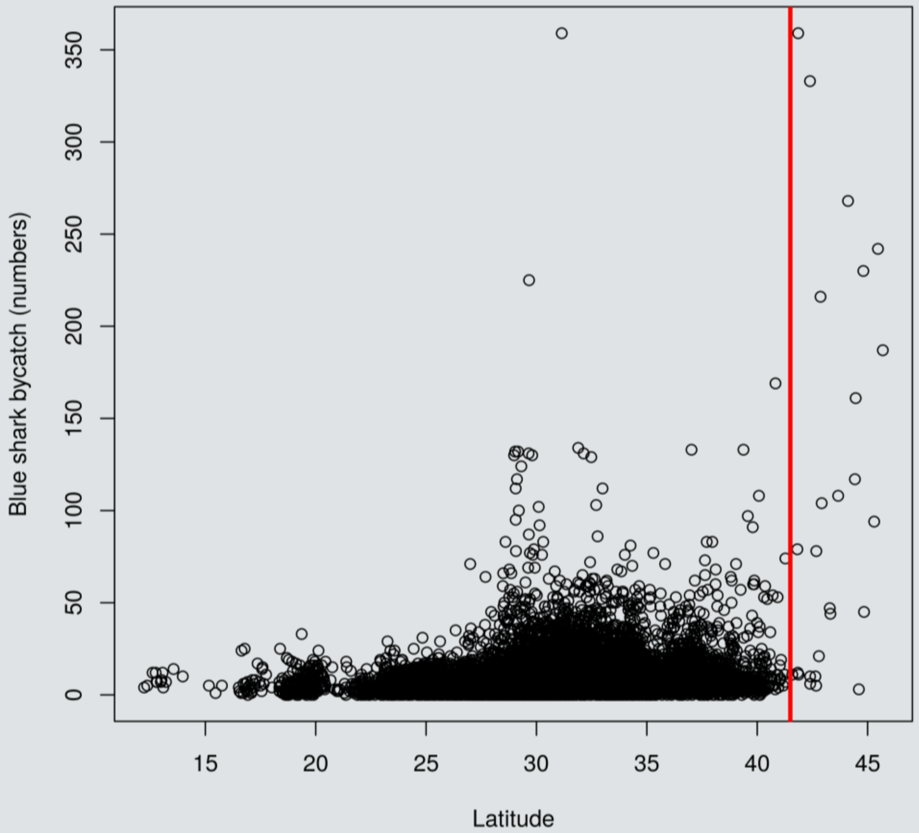
RF



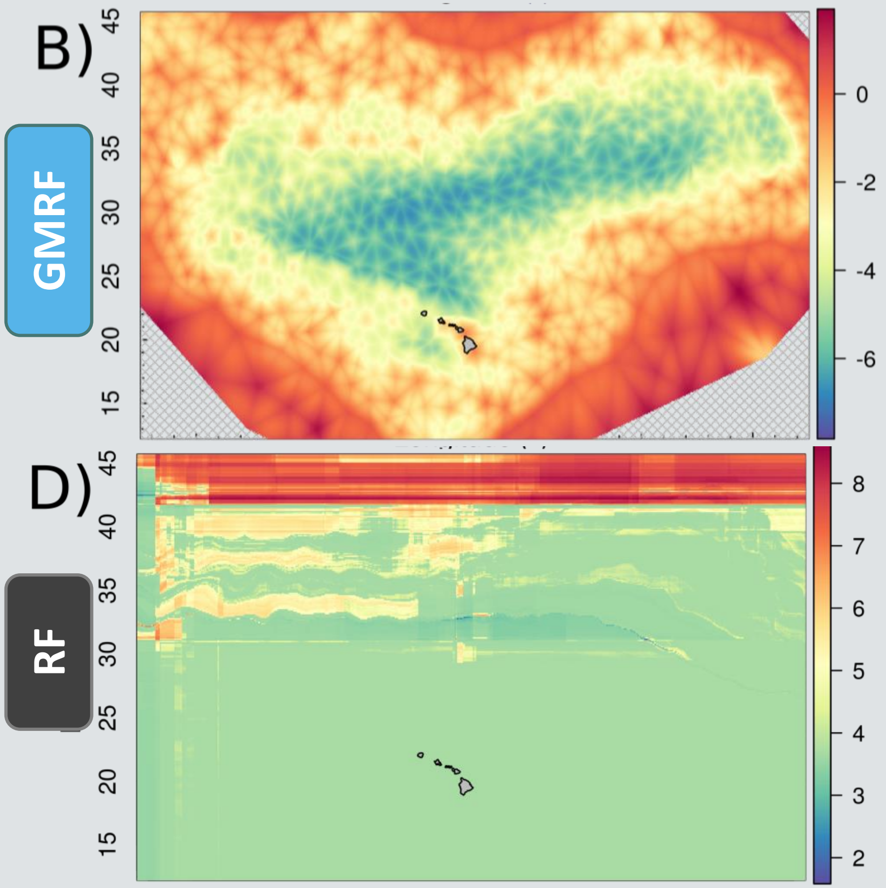
Wager et al. (2014)

Variance of predictions





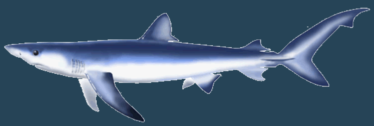
Variance



Wager et al. (2014)

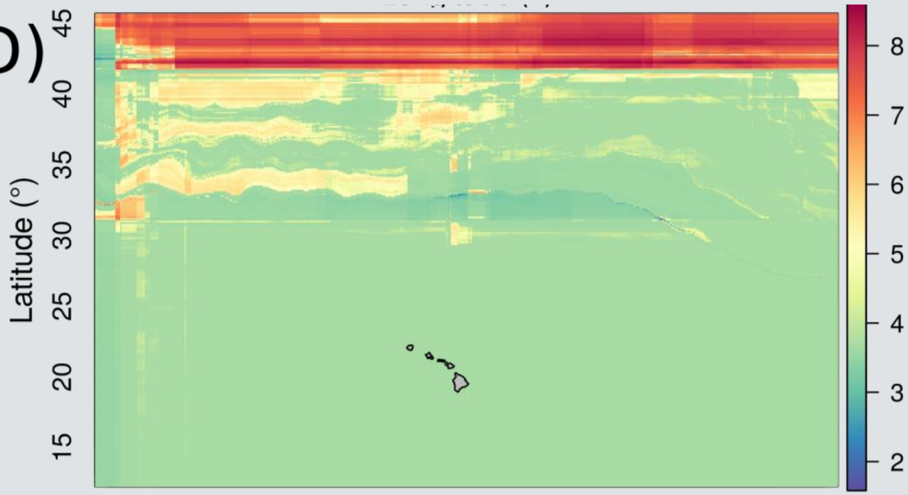
Variance of predictions

RF

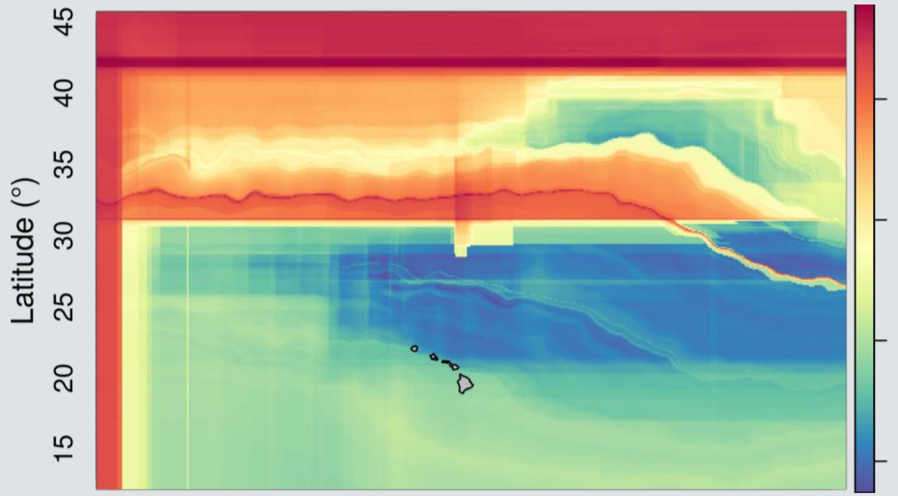


Variance

D)



Var(ind trees)



Non-parametric delta method /
“infinitesimal jackknife”

Wager et al. (2014)