Simon Dedman

gbm.auto

Automated Boosted Regression Tree software for spatial prediction of multiple life-history stock components

simondedman@gmail.com
simondedman.com

Background / specific case for work

- Spatial approaches to manage data-poor species
- Existing techniques often struggle / suboptimal
- Boosted Regression Trees (BRTs/GBMs) complicated but excellent performance:
  - Robust to poor/absent data
  - Can use abundance data
  - Unaffected by missing predictor values, outliers, multicollinearity
  - Can accommodate large numbers of explanatory variables without penalty
  - More robust predictions than GLMs and GAMs
  - Less variance (oversensitivity to noise leading to overfitting/imprecision)
  - Less bias (false assumptions in the algorithm leading to underfitting/inaccuracy)
  - Lower risk of misspecification
  - Ability to model complex interactions

Regression Tree models:

- Machine learning. No assumed relationship, model learns predictor-response relationships
- Uses algorithms to *partition* the *predictor space* into sections of the most *homogenous* response to predictors – blocks of reliable predictor-response relationship – carving out these blocks in binary splits at points along the predictors' ranges
- Predictors & split points calculated to minimise prediction error
- Not as accurate as GLMs/GAMs
- Bad at modelling smooth functions
- Very dependent on the sample data used, i.e. results aren't generalisable

Boosting

- "it is easier to find and average many rough rules of thumb, than to find a single, highly accurate prediction rule" (Elith *et al*. 2008)
- Finds one tree that best explains the predictors-response relationship, then
- Finds the tree that best explains the predictors-response relationship *of the residuals of the one-tree model* (which is a new tree with different values)
- Updates the model to incorporate the predictors-response relationship information gained from tree 1 plus tree 2
- Runs this new *2 tree* model on the data (choosing a different random testing chunk each time), producing new residuals. Makes new tree to test residuals, adds to model to make 3 tree model, runs 3 tree model. Repeats 1000s of times: remaining unaccounted-for deviance falls, until adding trees is unhelpful.

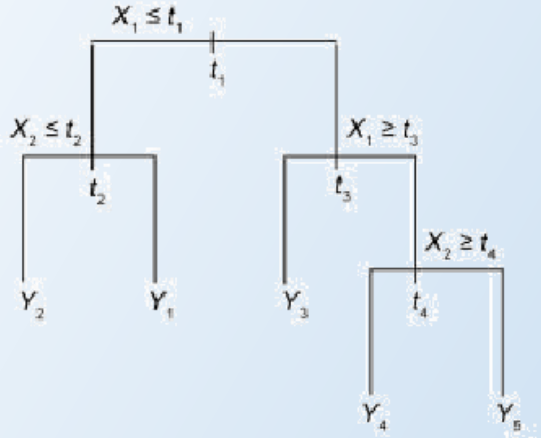
Journal of Animal Ecology

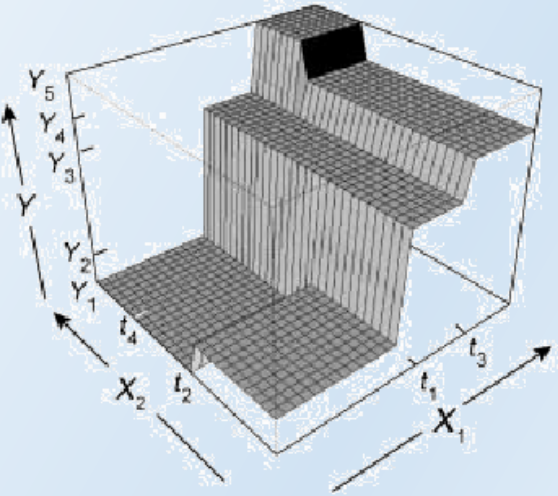
British Ecological Society

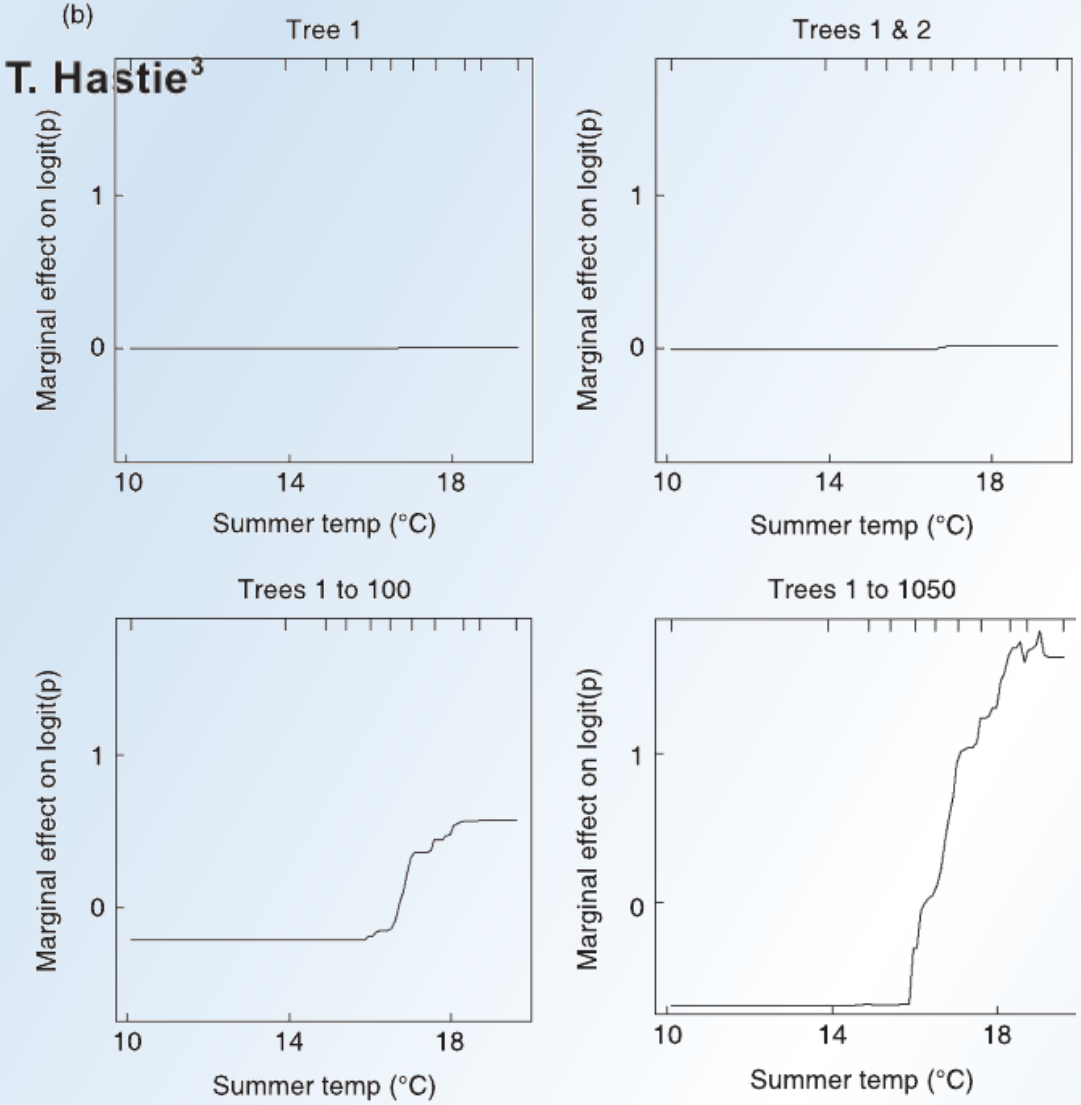
Journal of Animal Ecology 2008, **77**, 802–813 doi: 10.1111/j.1365-2656.2008.01390.x

A working guide to boosted regression trees

J. Elith[1]*, J. R. Leathwick[2] and T. Hastie[3]

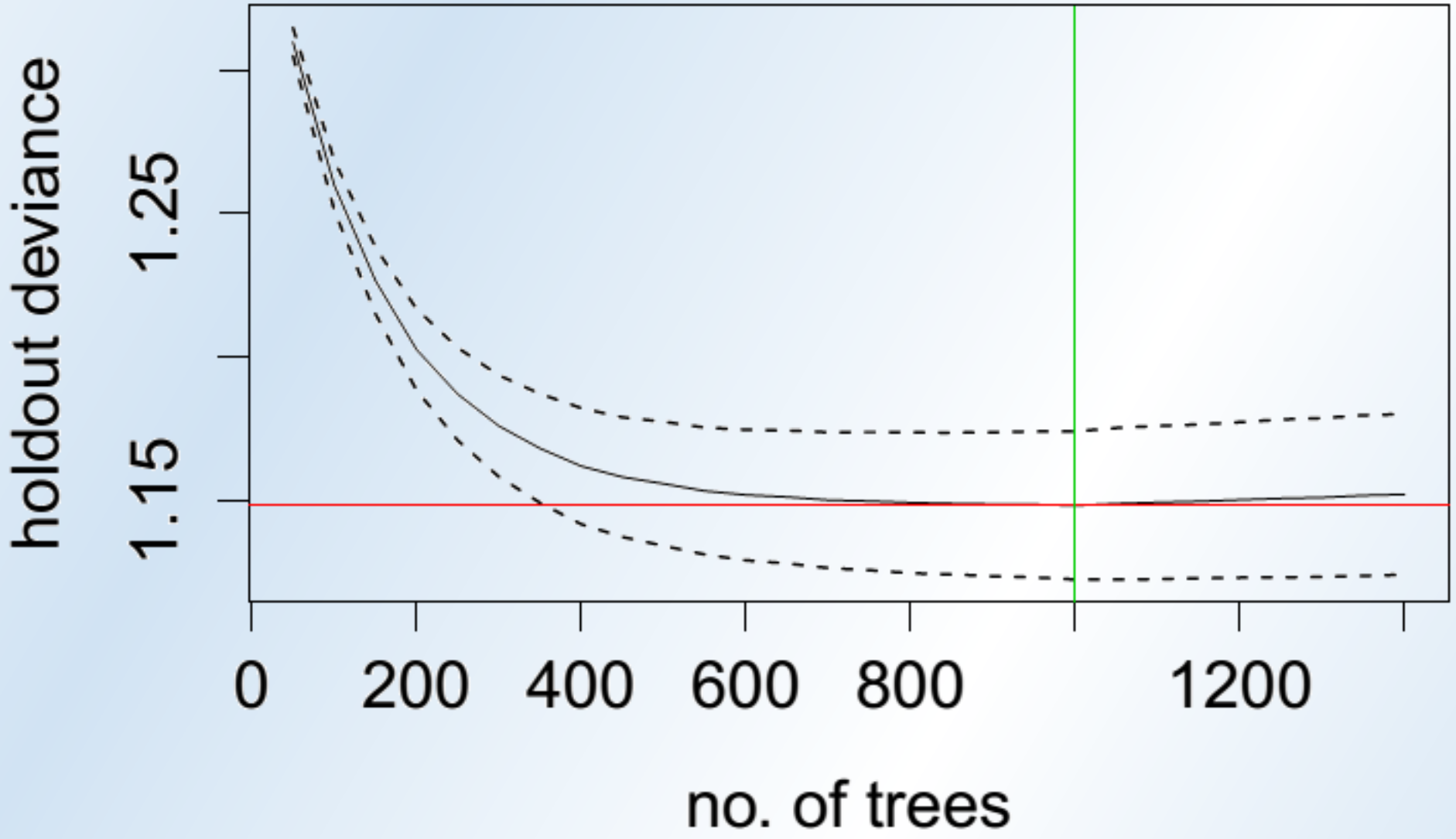






"Boosting is a numerical optimization technique for minimizing the loss function by adding, at each step, a new tree that best reduces (steps down the gradient of) the loss function. For BRT, the first regression tree is the one that, for the selected tree size, maximally reduces the loss function. For each following step, the focus is on the residuals: at the second step, a tree is fitted to the residuals of the first tree, and that second tree could contain quite different variables and split points compared with the first. The model is then updated to contain two trees (two terms), and the residuals from this two-term model are calculated, and so on."

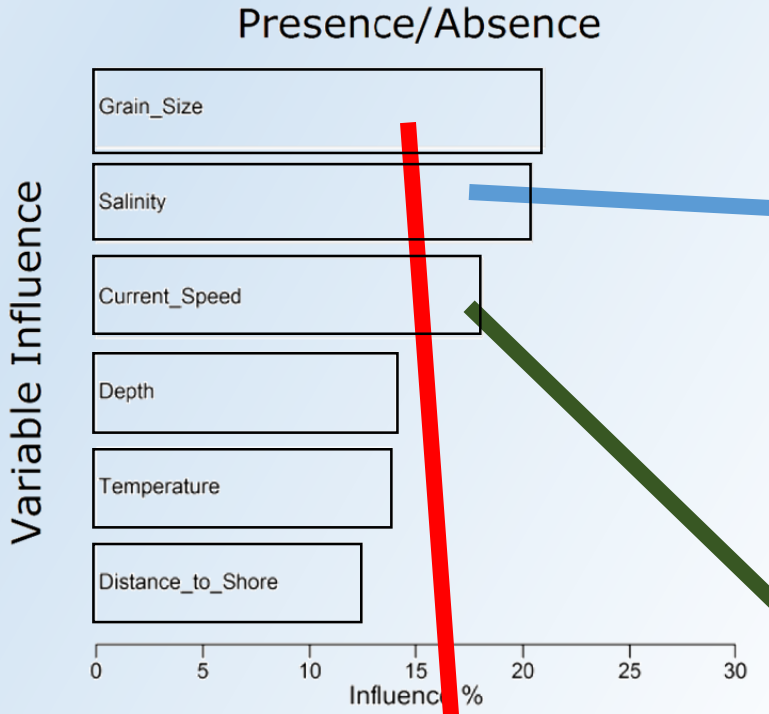
As the model incorporates more trees, the remaining unaccounted-for deviance falls, until the point where adding more trees adds unnecessary complexity and explains the predictors-response relationship LESS well. The code notes the number of trees which produce the lowest holdout deviance score, here 1000, and uses that model going forward.

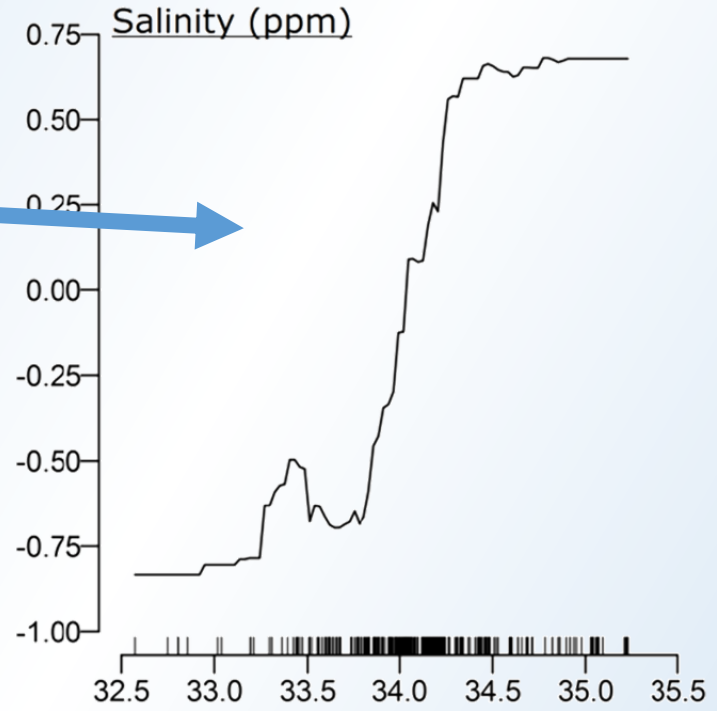


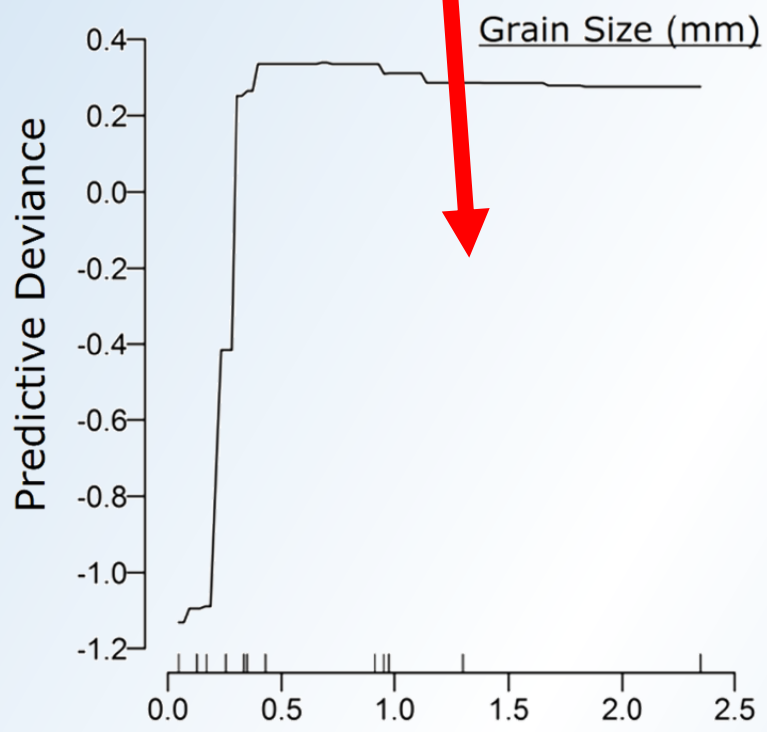
Relative contribution of each variable

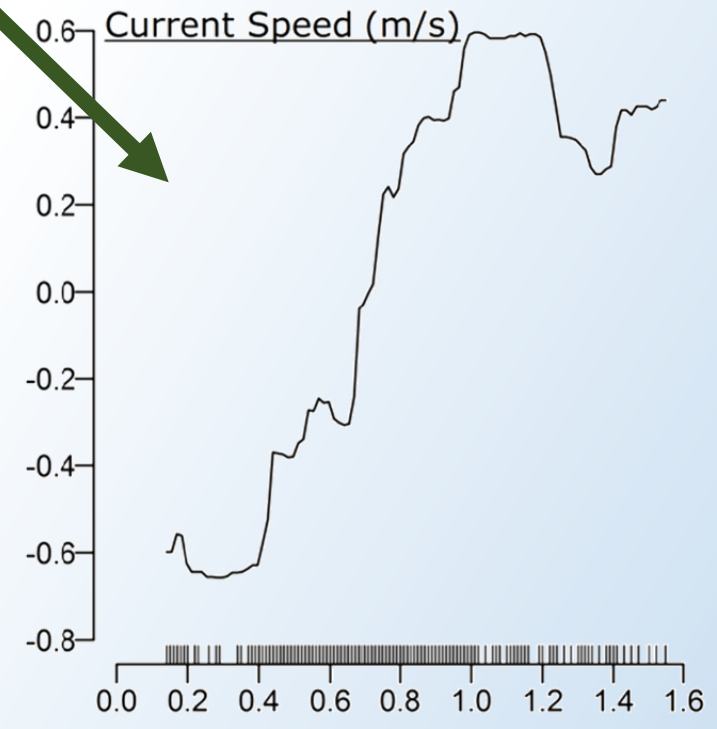
Predictor-response relationship for each variable

All of this complex information lives within the built model object. It's not completely a black box: you CAN force it to divulge its secrets, such as these figures, but its real value is using its knowledge to make predictions.





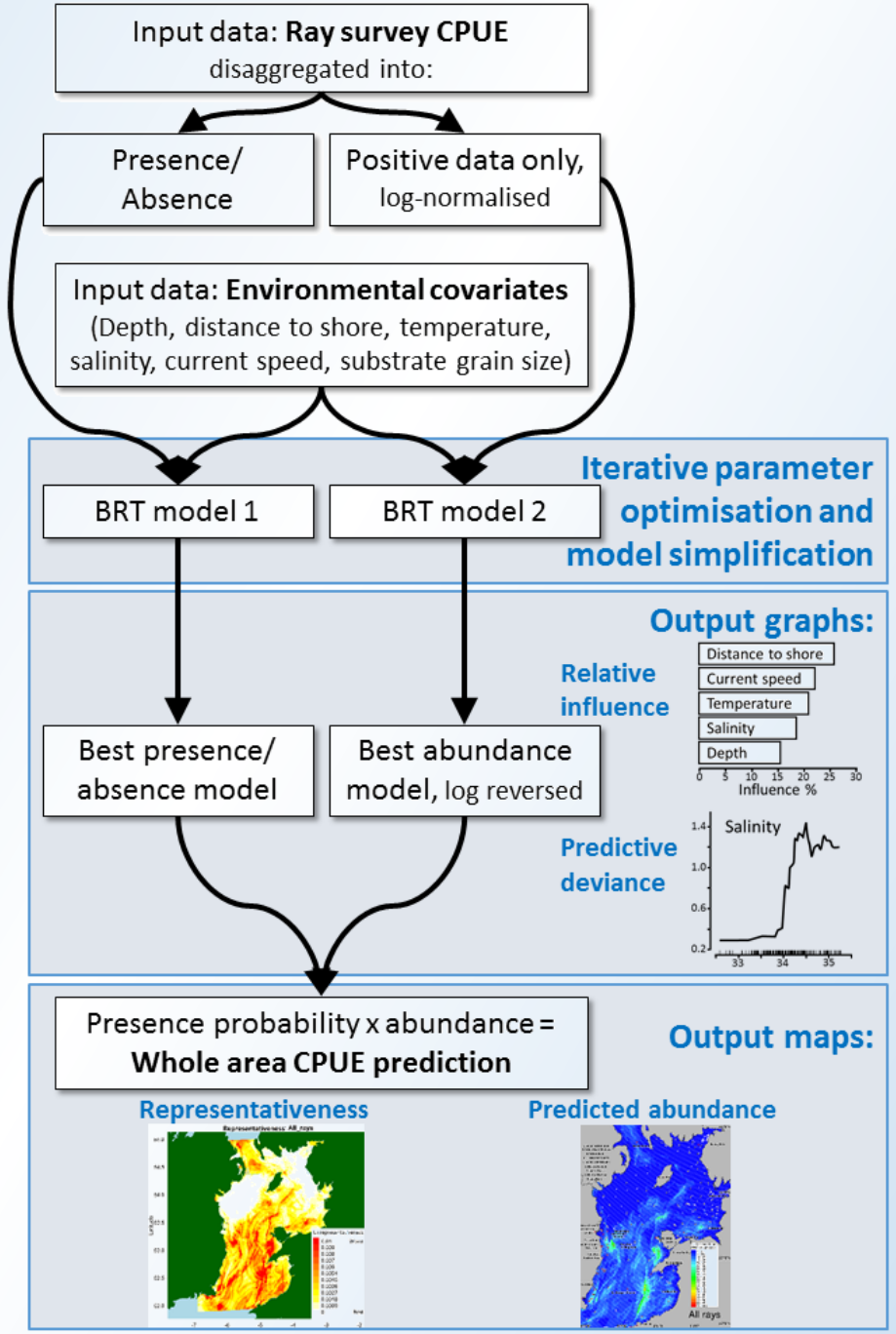


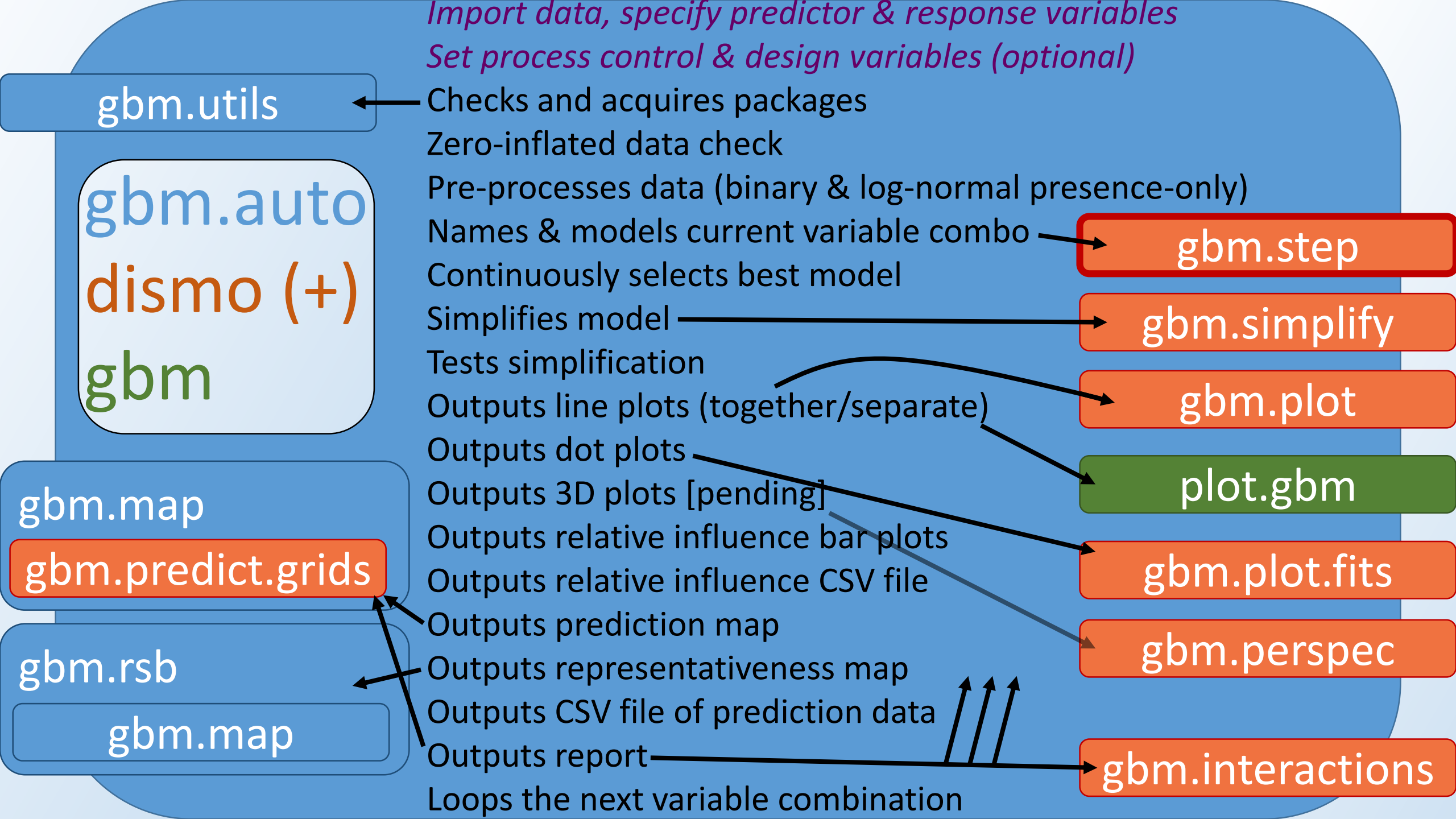


Work done to address specific need

Software suite in R that automates and greatly simplifies delta log-normal Boosted Regression Tree spatial modelling.

Powerful statistical modelling technique made accessible to potential users in the ecological and modelling communities.




Import data, specify predictor & response variables
Set process control & design variables (optional)
gbm.utils
Checks and acquires packages
Zero-inflated data check
Pre-processes data (binary & log-normal presence-only)
gbm.auto
dismo (+)
gbm
Names & models current variable combo
gbm.step
Continuously selects best model
Simplifies model
gbm.simplify
Tests simplification
Outputs line plots (together/separate)
gbm.plot
Outputs dot plots
plot.gbm
Outputs 3D plots [pending]
gbm.map
Outputs relative influence bar plots
gbm.predict.grids
gbm.plot.fits
Outputs relative influence CSV file
Outputs prediction map
gbm.perspec
gbm.rsb
Outputs representativeness map
gbm.map
Outputs CSV file of prediction data
Outputs report
gbm.interactions
Loops the next variable combination

Acquiring global coastlines with *gbm.basemap*

```
mybounds <- c(range(samples[,3]),range(samples[,2]))
gbm.basemap(bounds = mybounds)
```

Resolution 1, “coarse”
Resolution 5, “full”

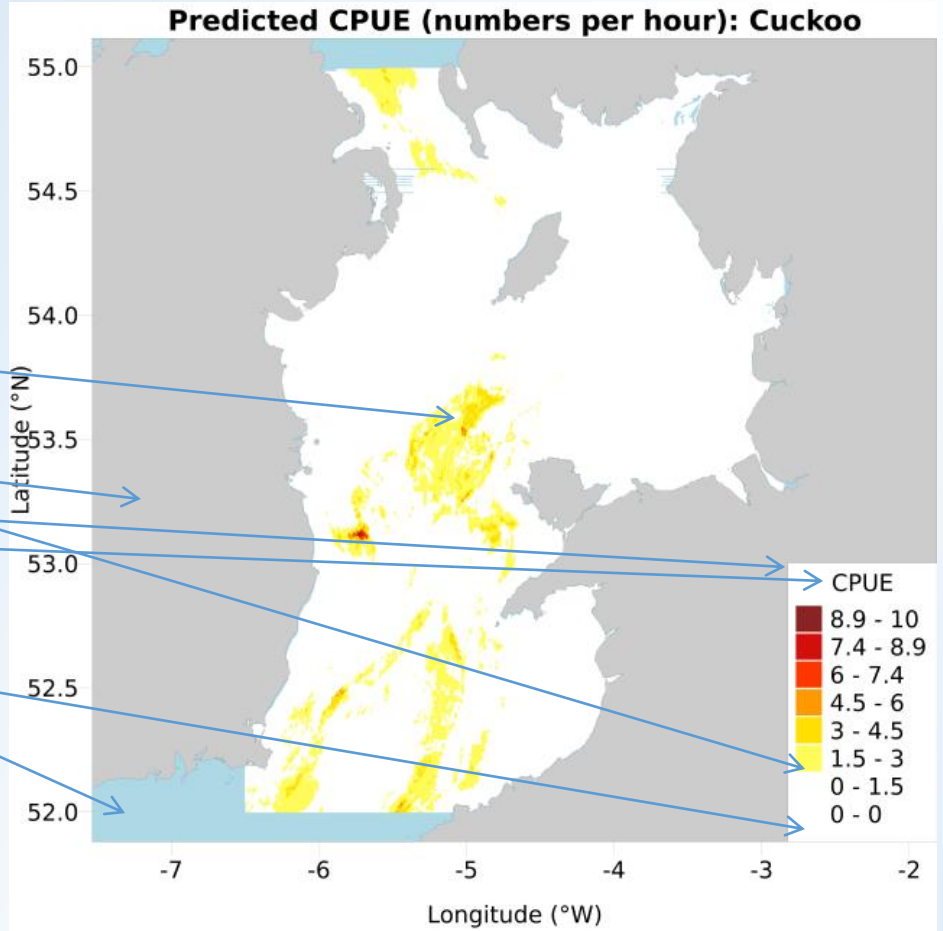
GBM = Gradient Boosting Machine / Generalised Boosted Models: EXACTLY the same thing as BRT Boosted Regressions Trees, but a different name.
And all the parameters have different name.
No idea why this issue exists.

Mapping with *gbm.map*

	A	B	C
1	LONGITUDE	LATITUDE	Abundance?
2	-6.45765	53.97035	0
3	-6.18835	52.2358	1
4	-6.1692	52.57755	0
5	-6.11505	52.7634	0
6	-6.115	52.7534	2
7	-6.1147	52.7636	3
8	-6.11315	53.6292	2
9	-6.1128	53.62605	0
10	-6.1128	54.0279	0
11	-6.11245	52.75085	2
12	-6.11245	52.7559	0

- Mapping function for gridded data
- calculates the cell size automatically
- allows user to alter most elements of the output

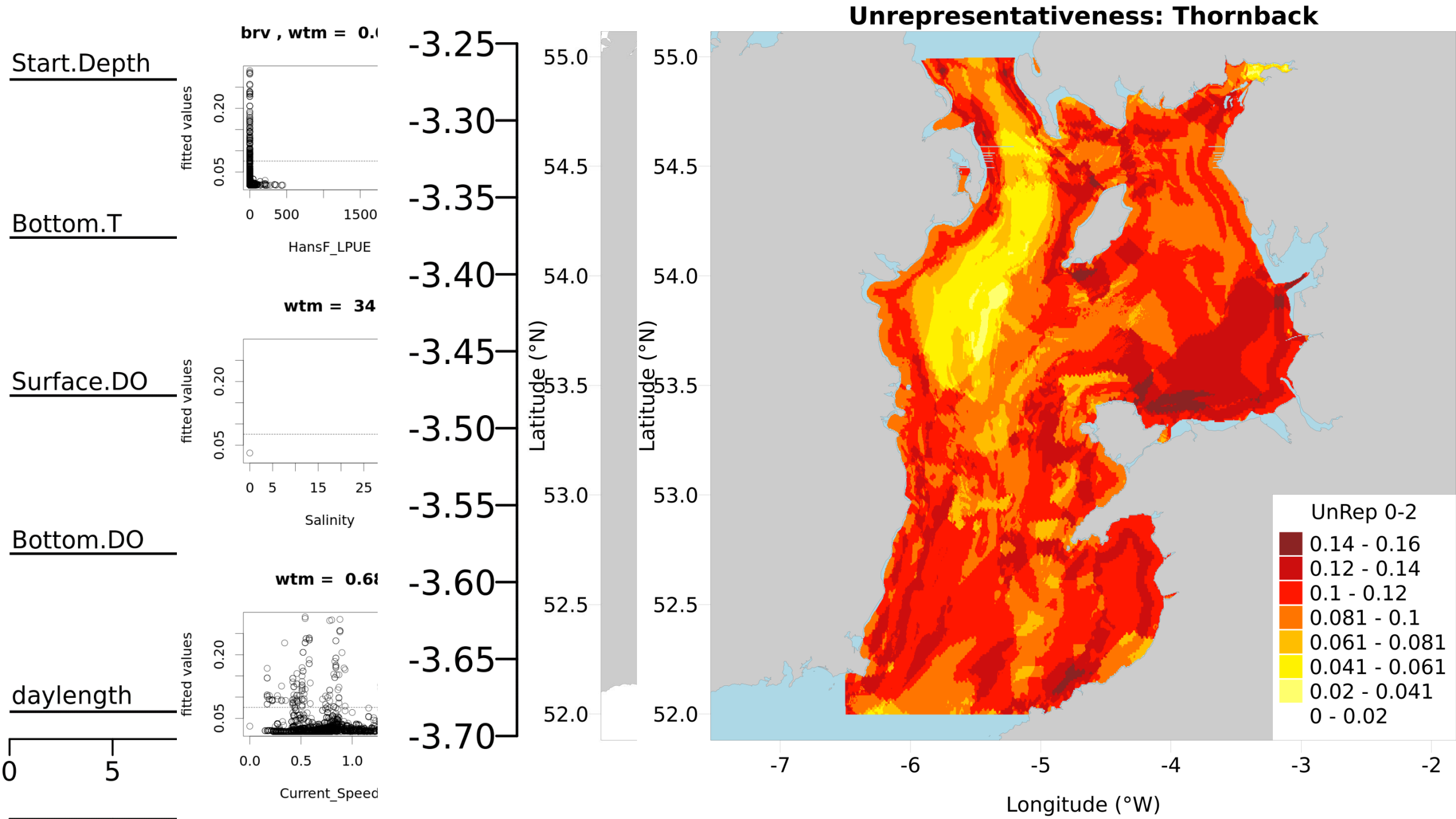
```
png(filename =  “ExampleMap.png")
par(mar = c(3.2,3,1.3,0))
gbm.map(x = grids[,1], y = grids[,2], z = grids[,3],
            species = “Cuckoo”,
            heatcolours
            colournumber
            landcol
            mapback
            legendloc
            legendtitle
            lejback
            etc)
dev.off()
```



Abundance predictions with *gbm.auto*

```
samples <- read.csv(“samples.csv")
grids <- read.csv(“grids.csv")
gbm.auto(samples = samples, grids = grids, expvar = 4:6, resvar = 3)
```

- Uses gbm.basemap, gbm.map, gbm.rsb and various other functions
- Allows the user to specify which data distribution to use
- can check for zero-inflation and transform data to use the delta-lognormal model on long-tailed zero-inflated data
- automatically loops through the user-set combinations of parameters and multiple response variables


Start.Depth
Bottom.T
Surface.DO
Bottom.DO
daylength
0
5
brv , wtm = 0.
fitted values
0.05
0.20
0
500
1500
HansF_LPUE
wtm = 34
fitted values
0.05
0.20
0
5
15
25
Salinity
wtm = 0.68
fitted values
0.05
0.20
0.0
0.5
1.0
Current_Speed
-3.25
-3.30
-3.35
-3.40
-3.45
-3.50
-3.55
-3.60
-3.65
-3.70
Latitude (°N)
55.0
54.5
54.0
53.5
53.0
52.5
52.0
Unrepresentativeness: Thornback
Latitude (°N)
55.0
54.5
54.0
53.5
53.0
52.5
52.0
-7
-6
-5
-4
-3
-2
Longitude (°W)
UnRep 0-2
0.14 - 0.16
0.12 - 0.14
0.1 - 0.12
0.081 - 0.1
0.061 - 0.081
0.041 - 0.061
0.02 - 0.041
0 - 0.02

Explanatory Variables	Response Variables	Zero Inflated?	Bin_BRT.tc2.lr0.001.bf0.5	Best Binary BRT
Year	F.blacknose	TRUE	trees: 1400	Model combo: Bin_BRT.tc2.lr0.001.bf0.5
Season			Training Data Correlation: 0.588134932326192	Model CV score: 0.588134932326192
Lon			CV Mean Deviance: 1.08280416606036	Training data AUC score: 0.8827
Surface.T			CV Deviance SE: 0.0478621899131645	CV AUC score: 0.6778
Bottom.T			CV Mean Correlation: 0.27417021183714	CV AUC se: 0.0591583505892824
Surface.DO			CV Correlation SE: 0.0886817493754063	
Bottom.DO				
Start.Depth				

Bin_BRT_simp predictors kept (ordered)	Bin_BRT_simp predictors dropped	Simplified Binary BRT stats
Season	Surface.DO	trees: 2600
Depth.Bin	Start.Depth	Training Data Correlation: 0.564892280252173
Year		CV Mean Deviance: 1.04457745738497
Surface.T		CV Deviance SE: 0.0497216516054207
Bottom.T		CV Mean Correlation: 0.305782672406723
daylength		CV Correlation SE: 0.0775726580103382
Bottom.DO		
Lon		

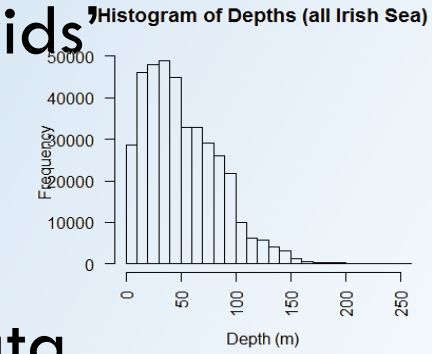
Best Binary BRT variables	Relative Influence (Bin)	Biggest Interactions (Bin)
Start.Depth	35.54444975	Start.Depth and Lon. Size: 1.18
Surface.DO	16.89503265	Bottom.DO and Surface.DO. Size: 0.25
Lon	15.41003404	
daylength	9.070503635	
Bottom.DO	7.959983717	
Bottom.T	7.63123564	
Surface.T	3.550914543	
Year	2.09227579	
Depth.Bin	1.417916858	
Season	0.427653375	

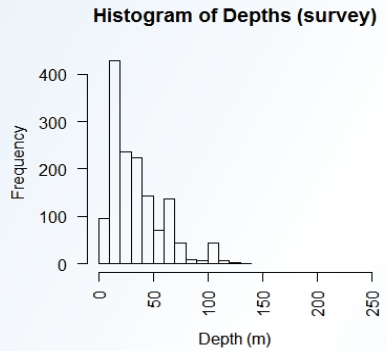
The area under the ROC curve can be integrated and interpreted as an Area Under the Curve (AUC) value that has a range from 0.5 to 1. Using this metric, a value of one indicates perfect discrimination of probabilities between presence and absence samples and a value of 0.5 indicates that model discrimination is no better than random. While models with AUC values greater than 0.6 are considered useful (Parisien and Moritz 2009), values greater than 0.8 are considered very good, and above greater than 0.9 excellent (Lane et al. 2009).

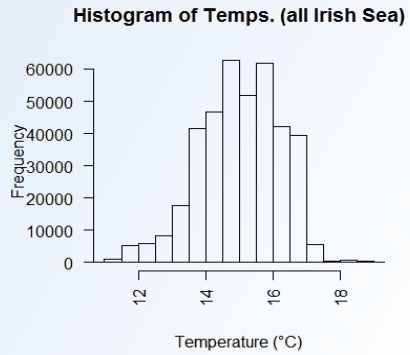
Visual assessment of data quality and representativeness with *gbm.rsb*

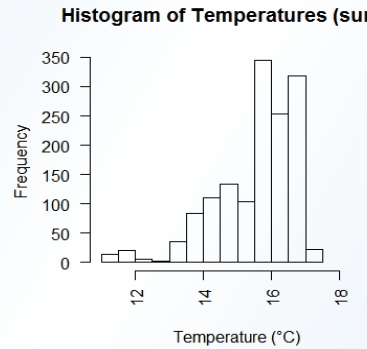
Representativeness Surface Builder

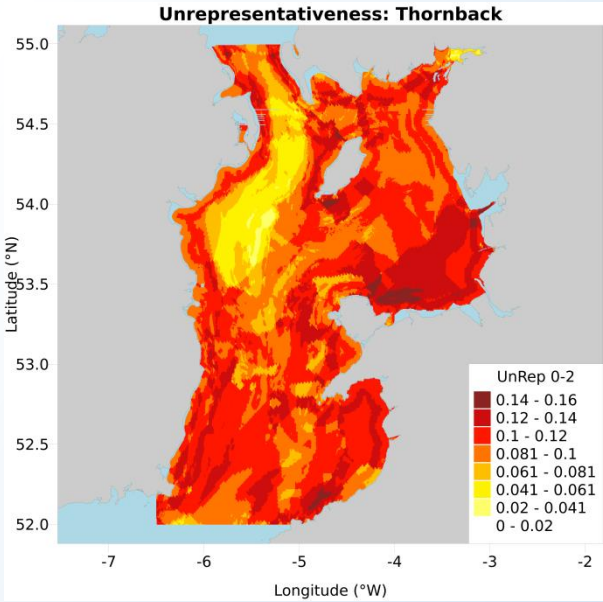
- Compares frequency distribution of the explanatory variables from the 'grids' data with those from the 'samples' data
- Differences summed into a score indicating how well the samples data captures that variable's full range
- Calculated for every cell in 'grids'
- Exported to csv & mapped with *gbm.map*
- Higher values = poor coverage = be more cautious with conclusions at that point.







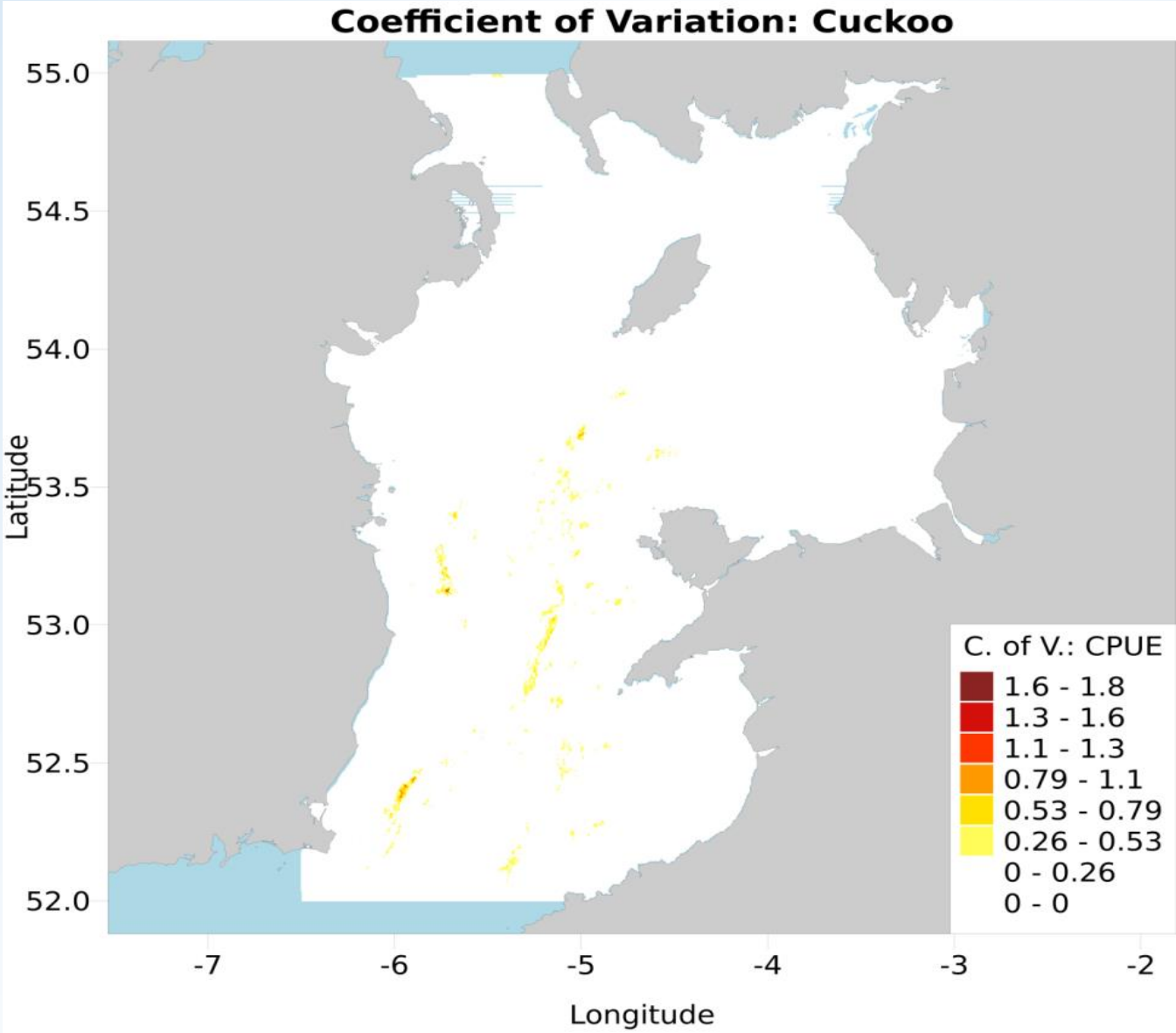


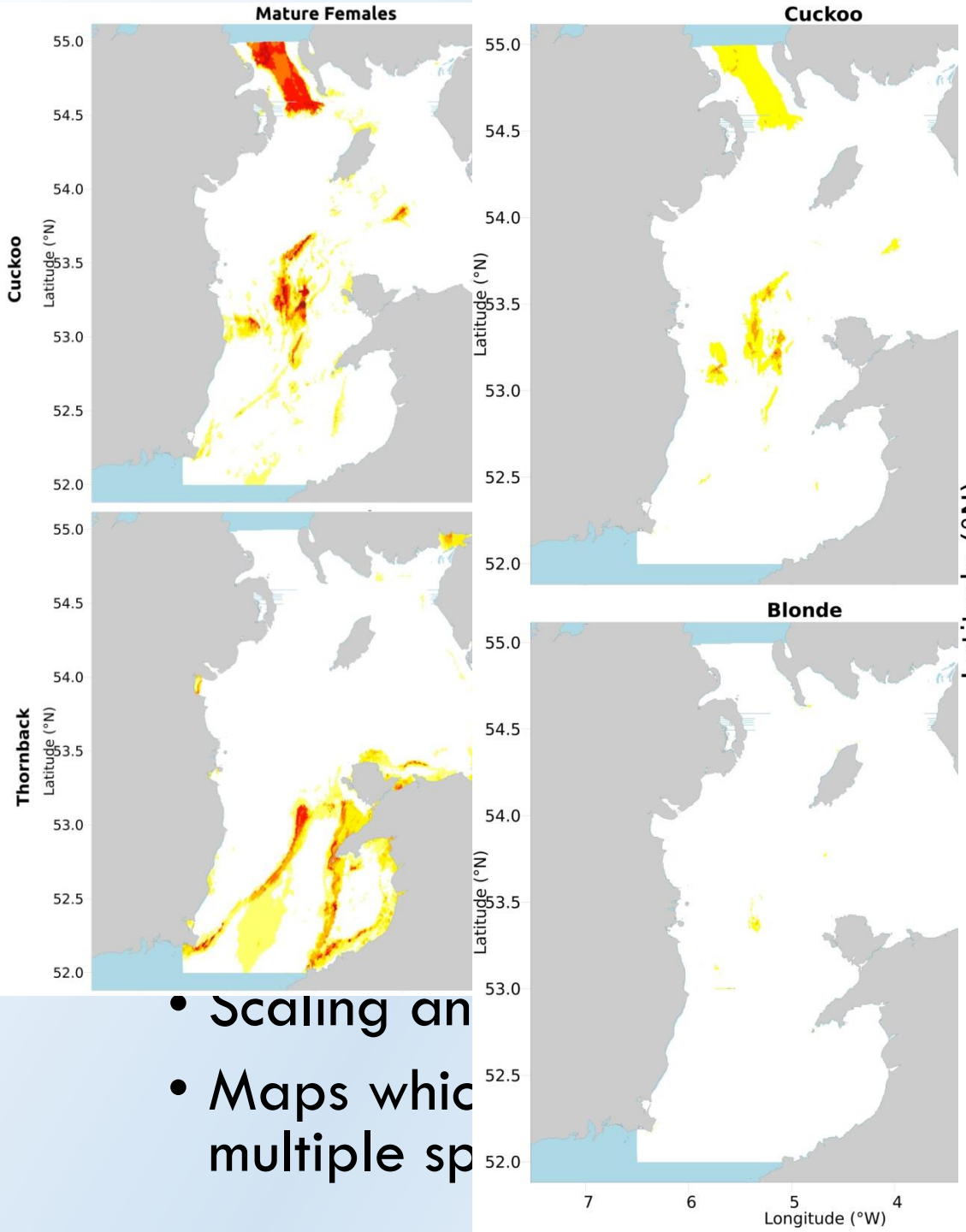


gbm.rsb(samples, grids, expvarnames, gridslat, gridslon)

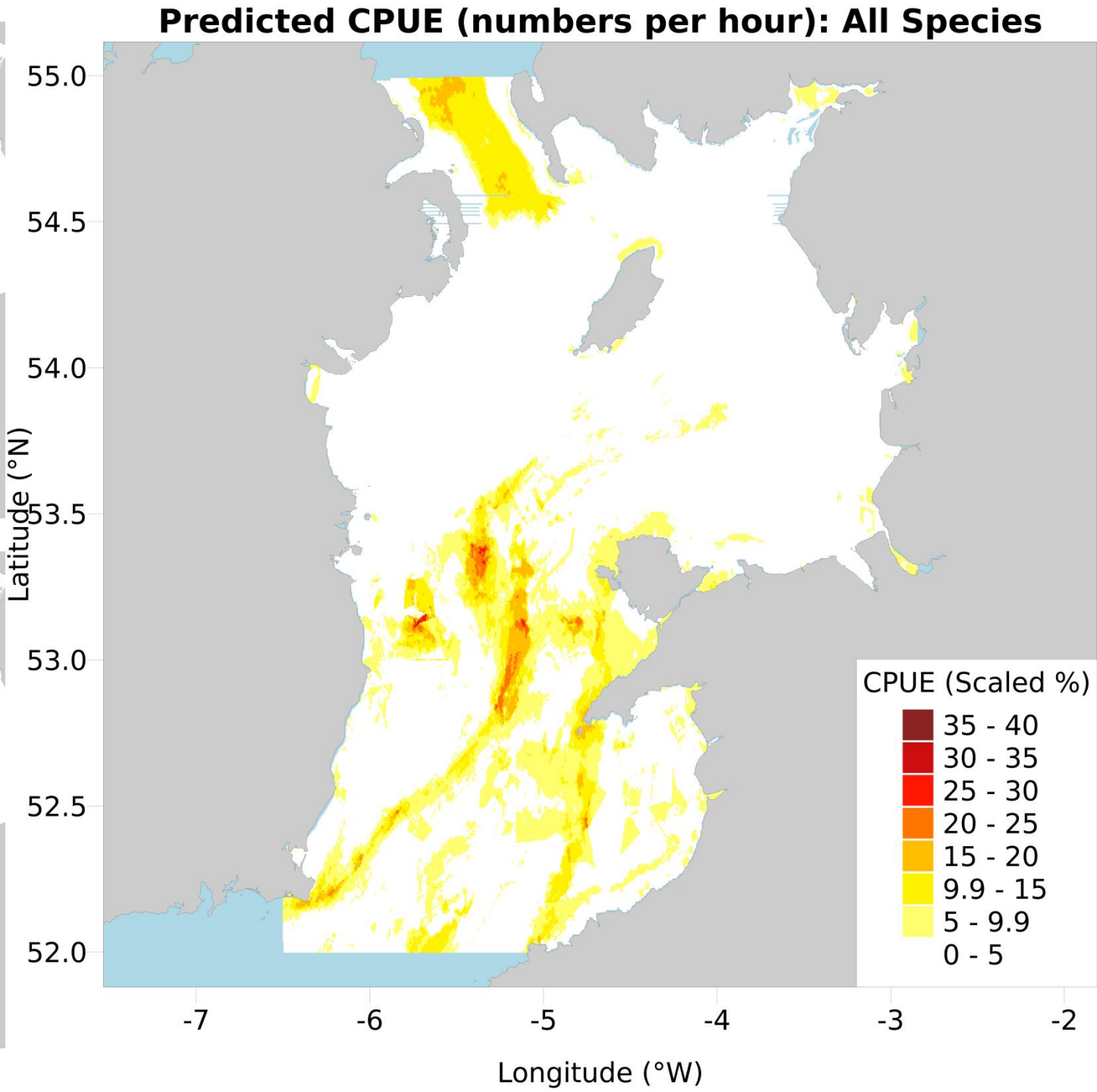
Calculating the coefficient of variation of predicted abundance with *gbm.loop*

- Repeats *gbm.auto* run a user-specified number of times
- Calculates and plots the minimum, average, maximum, and variance of the variable influence values (bar plot data)
- Calculates and plots the minimum, average, and maximum partial dependence values (line plot data)
- Calculates coefficient of variation for predicted abundance map.
- Produces map and csv files





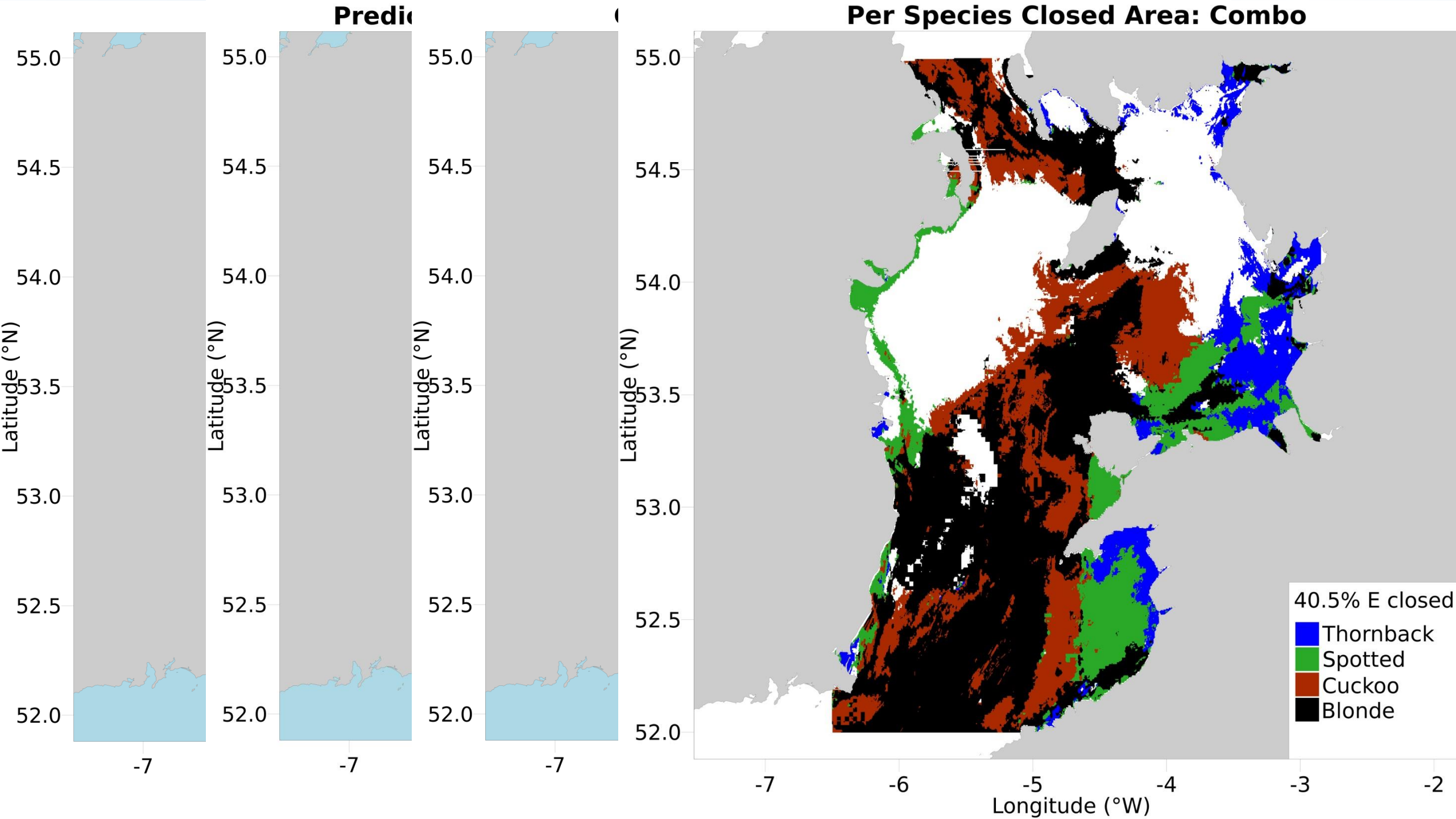
- Scaling an
- Maps whic
  multiple sp



Generating MPAs with *gbm.valuemap*

```
gbm.valuemap(dbase = mydata, goodcols = c(5,3,6,4),  badcols = 7, conservecol = 8, HRMSY = c(0.14,0.08,0.08,0.15))
```

- Predictive maps only addresses half the problem.
- Conservation plans are prioritisations: must consider socioeconomic metrics e.g. fishing effort
- Need biologically-derived MPA candidates. Maximum Sustainable Yield (MSY) principle of *escapement biomass:* percentage to retain annually to conserve the stock, Harvest Rate at MSY (HR_{MSY}).
- Predicted abundance map of rays Vs map of fishing effort = areas to preferentially conserve, and areas to avoid closing to minimise effort displacement.
- Cumulatively add cells sorted from most to least preferable to close until you have an MPA big enough to protect the most conservationally valuable species' HR_{MSY} ("species 1")
- Do the same with Species 2, but with Species 1's MPA already in place, i.e. you just grow Species 1's MPA until it protects Species 2. Repeat for all species.
- Instead of 'abundance Vs effort' prioritisation sort, can sort by effort only, abundance only, or conservation map areas from gbm.cons


Predi
Per Species Closed Area: Combo
55.0
54.5
54.0
53.5
53.0
52.5
52.0
Latitude (°N)
-7
-6
-5
-4
-3
-2
Longitude (°W)
40.5% E closed
Thornback
Spotted
Cuckoo
Blonde

Pre-run parameter scoping with *gbm.bfcheck*

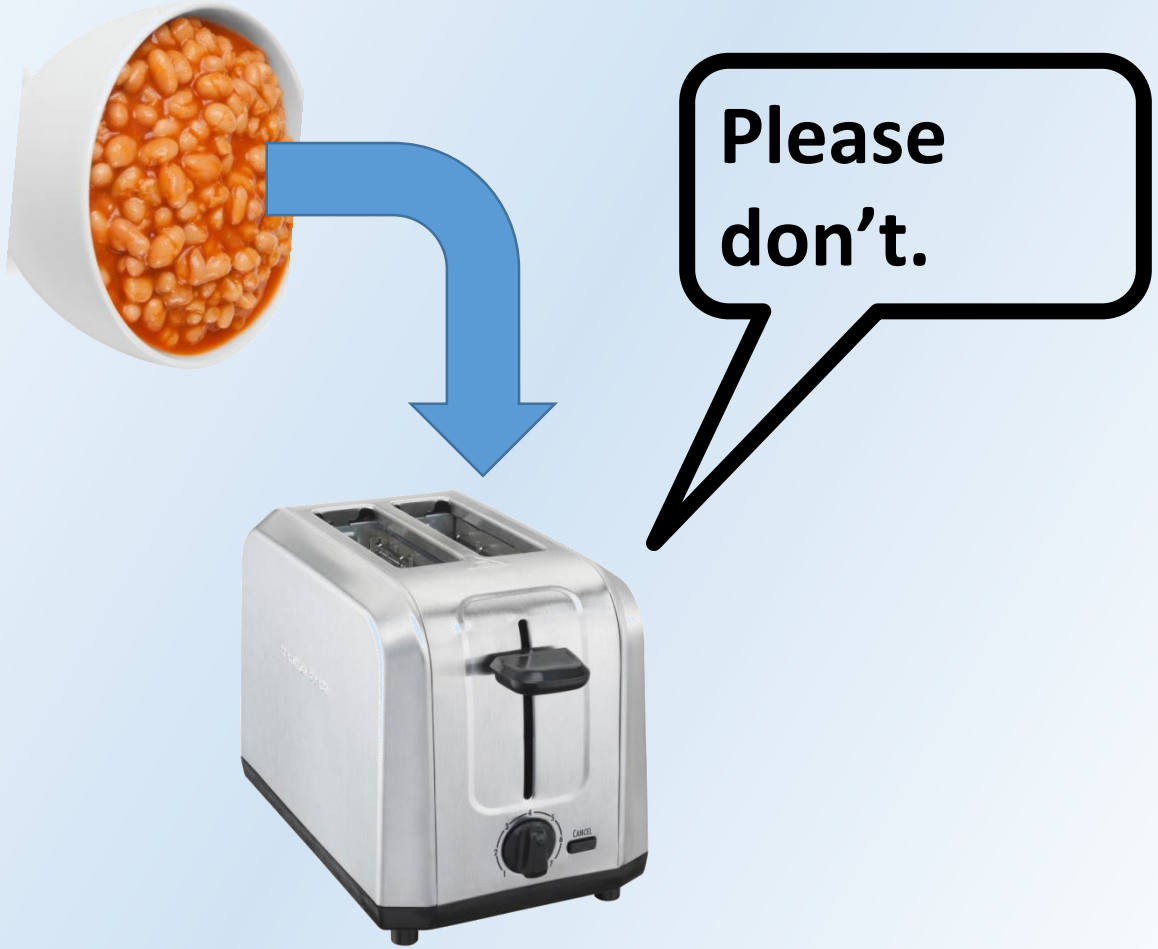
```
gbm.bfcheck(samples = samples, resvar = 3)
```

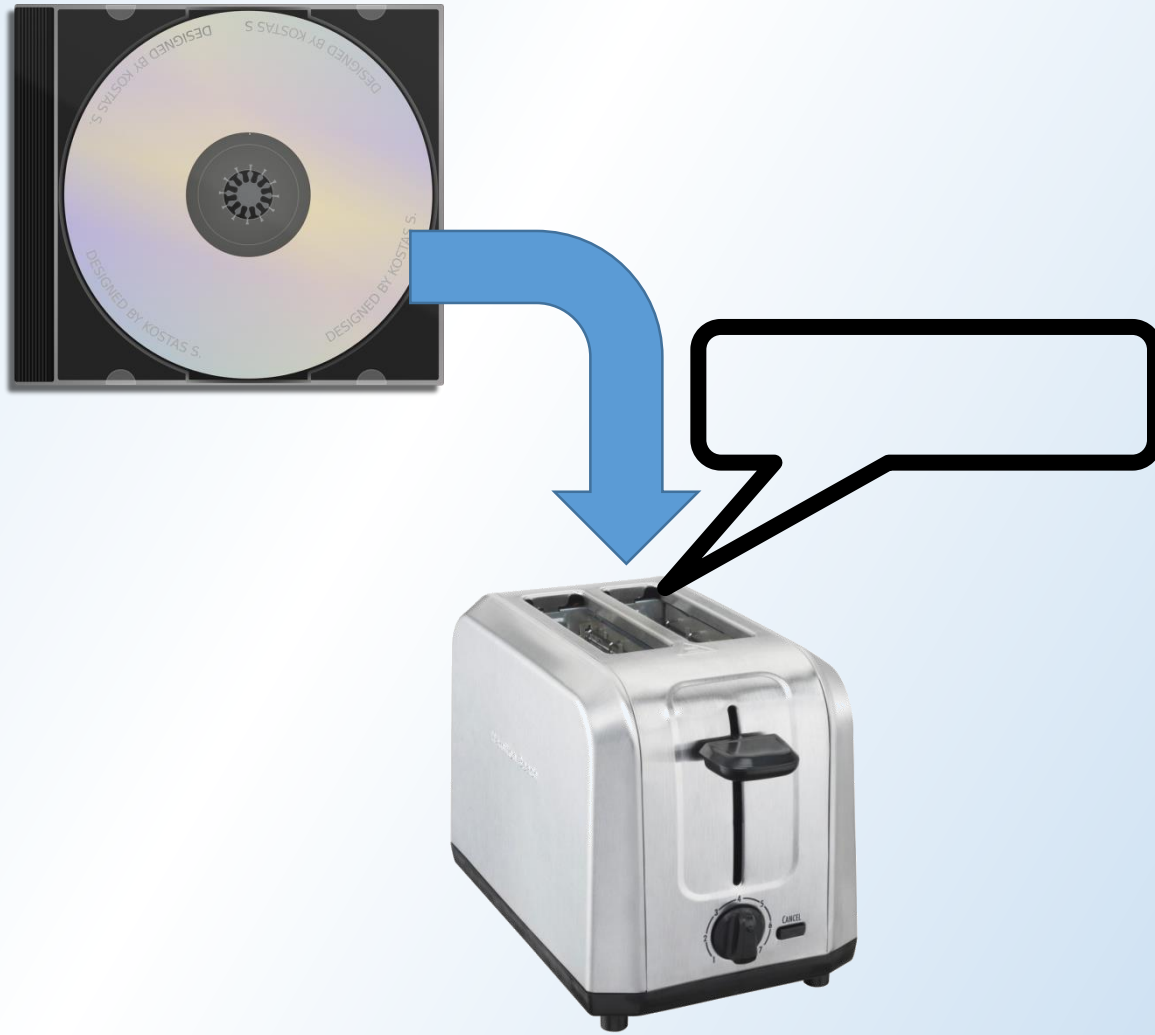
- Calculates the minimum binary and Gaussian BRT bag fraction sizes
- Users can check and optimise BFs before starting *gbm.auto* runs

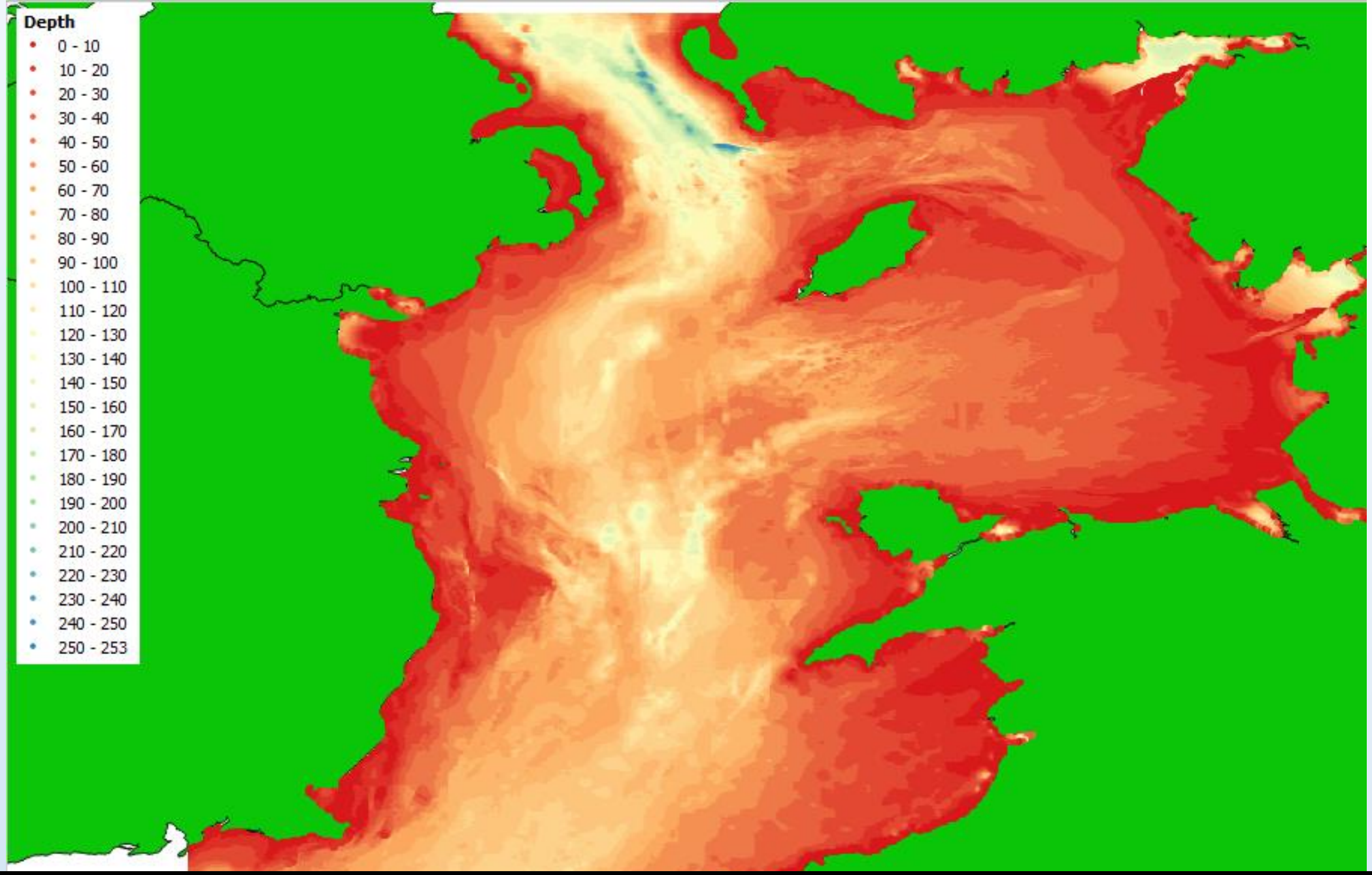
Conclusions

- Easily-usable and feature-rich resource.
- Data-poor or rich; single or multiple species or subsets.
- Users can easily produce predicted abundance maps, explanatory variable diagnoses, conservation priority area maps and area closure proposals, with little work or prior knowledge required.
- Facilitate and expedite conservation of data-poor species using MPAs that balance competing priorities with the full engagement of stakeholders.
- Customisability means users can reduce analyses to the essentials they require.
- Users can quickly generate high quality outputs for presentations and journals, without lengthy/repeated formatting.
- Output maps and plots can drive collaborative MPA siting discussions with stakeholders and fisheries managers.

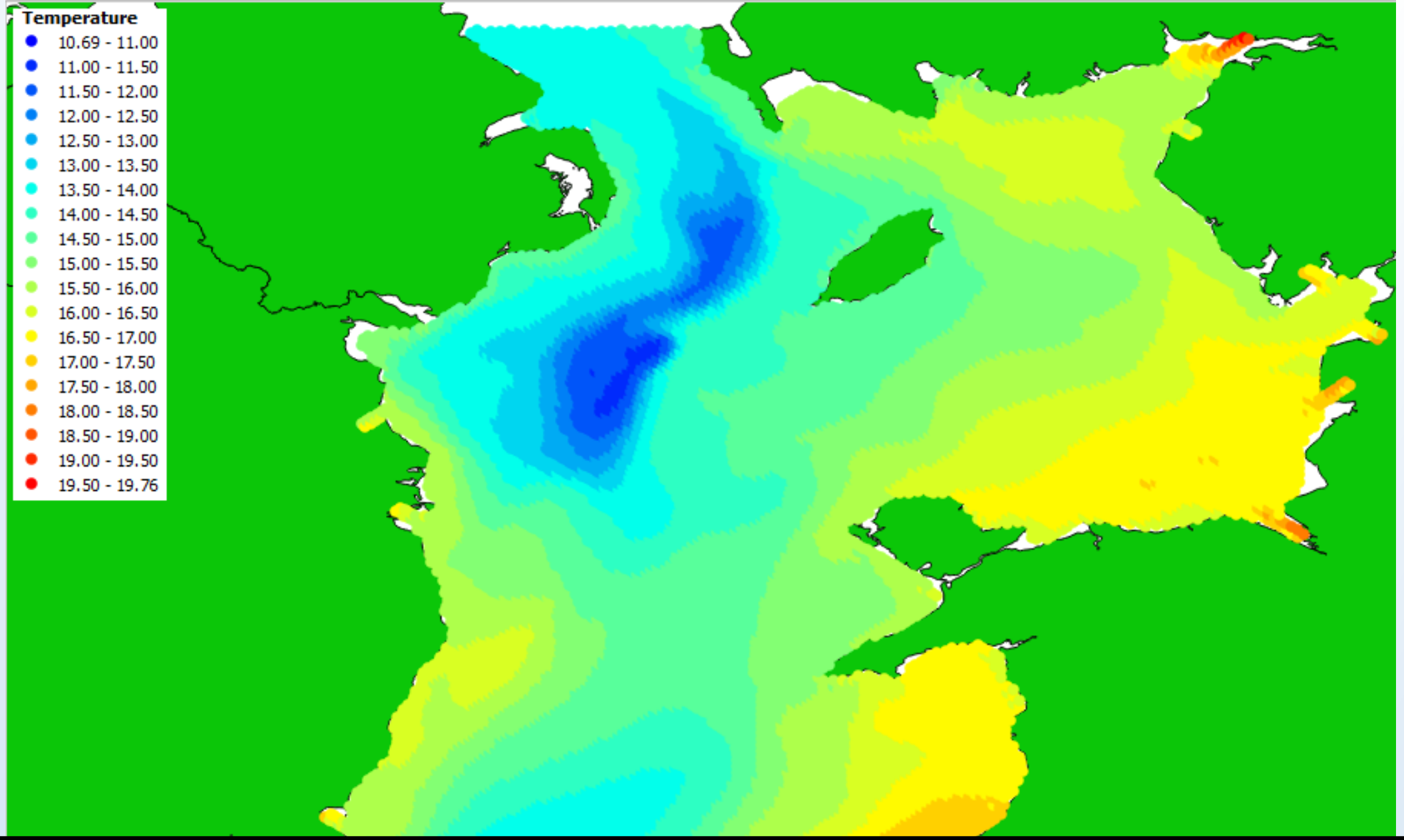
“Prey Mr Babbage, if you put into the machine wrong figures, will right answers come out?”



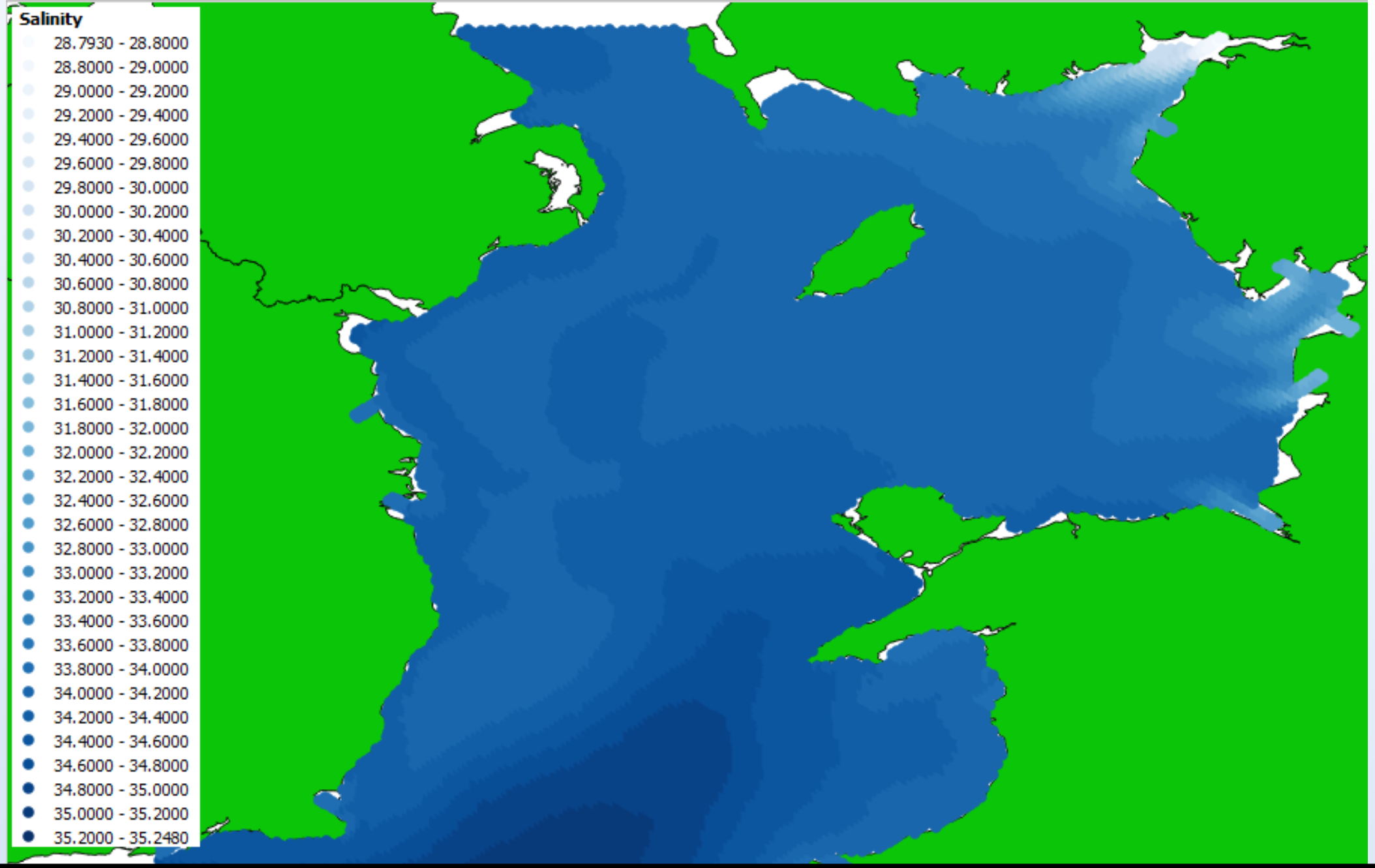




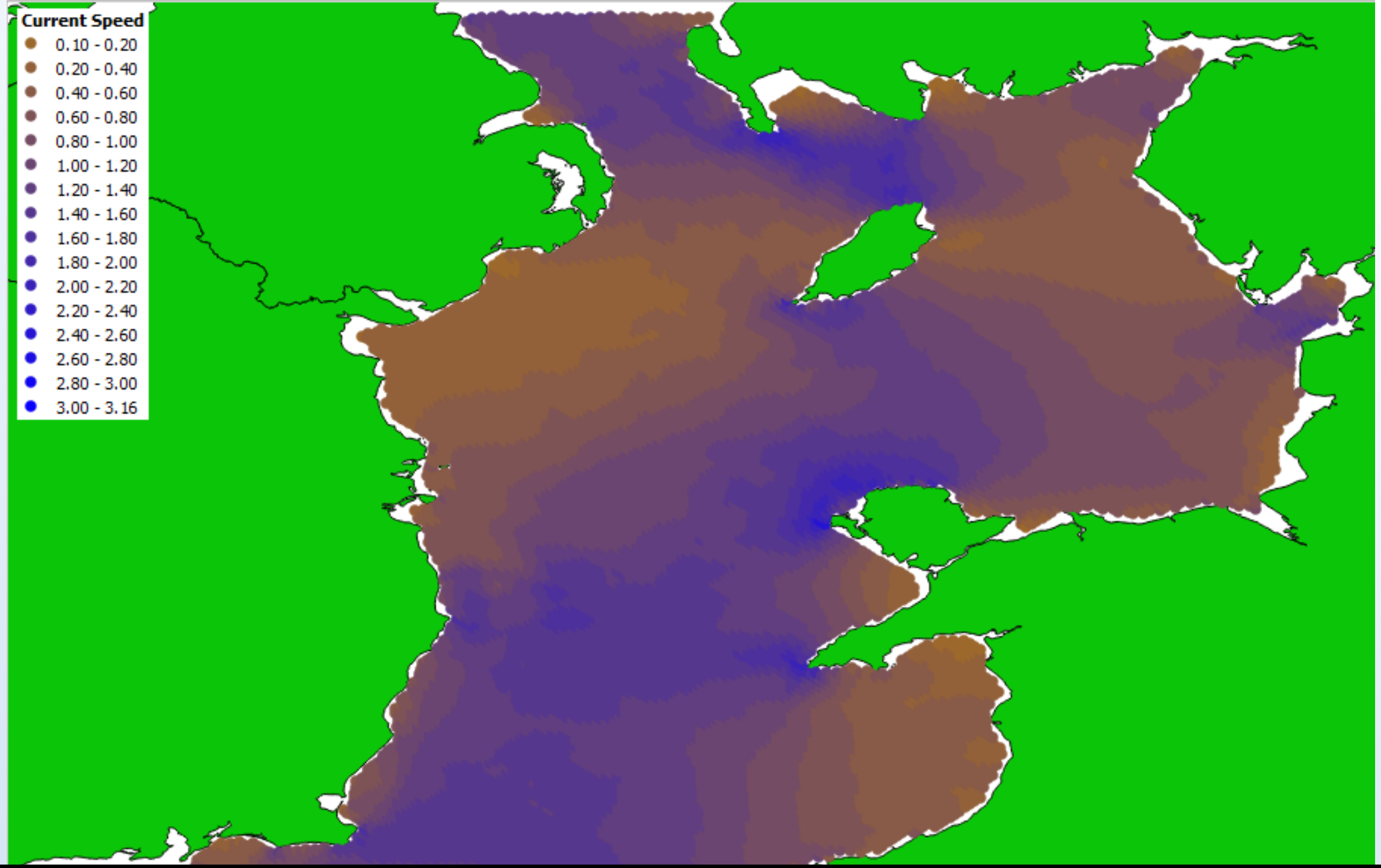
Depth: 391,568 275x455m grids, European Marine Observation and Data Network



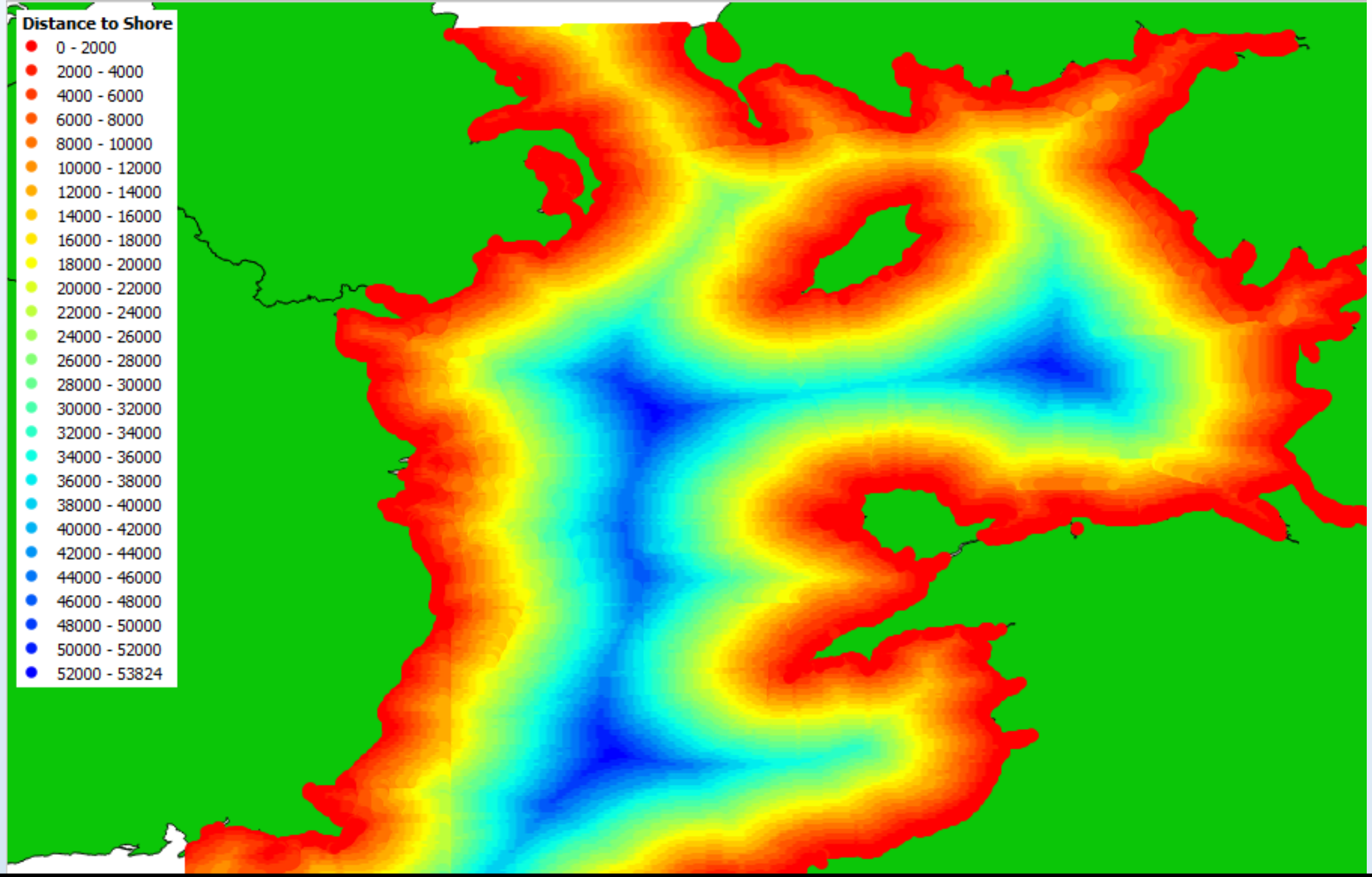
Average monthly sea bottom temperature 2010-2012: 22506 1185x1680m grids, Marine Institute



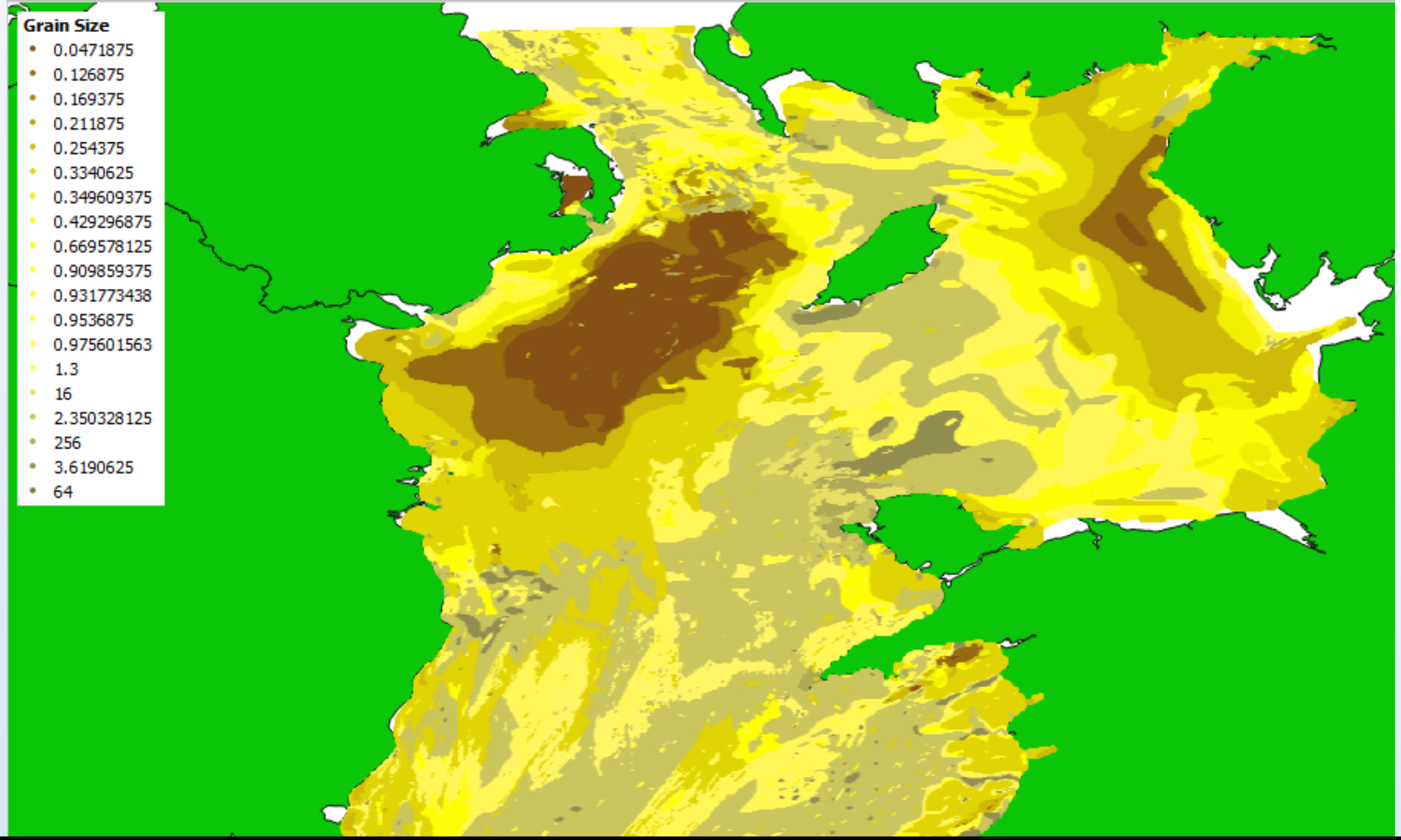
Average monthly sea bottom salinity 2010-2012: 22506 1185x1680m grids, MI



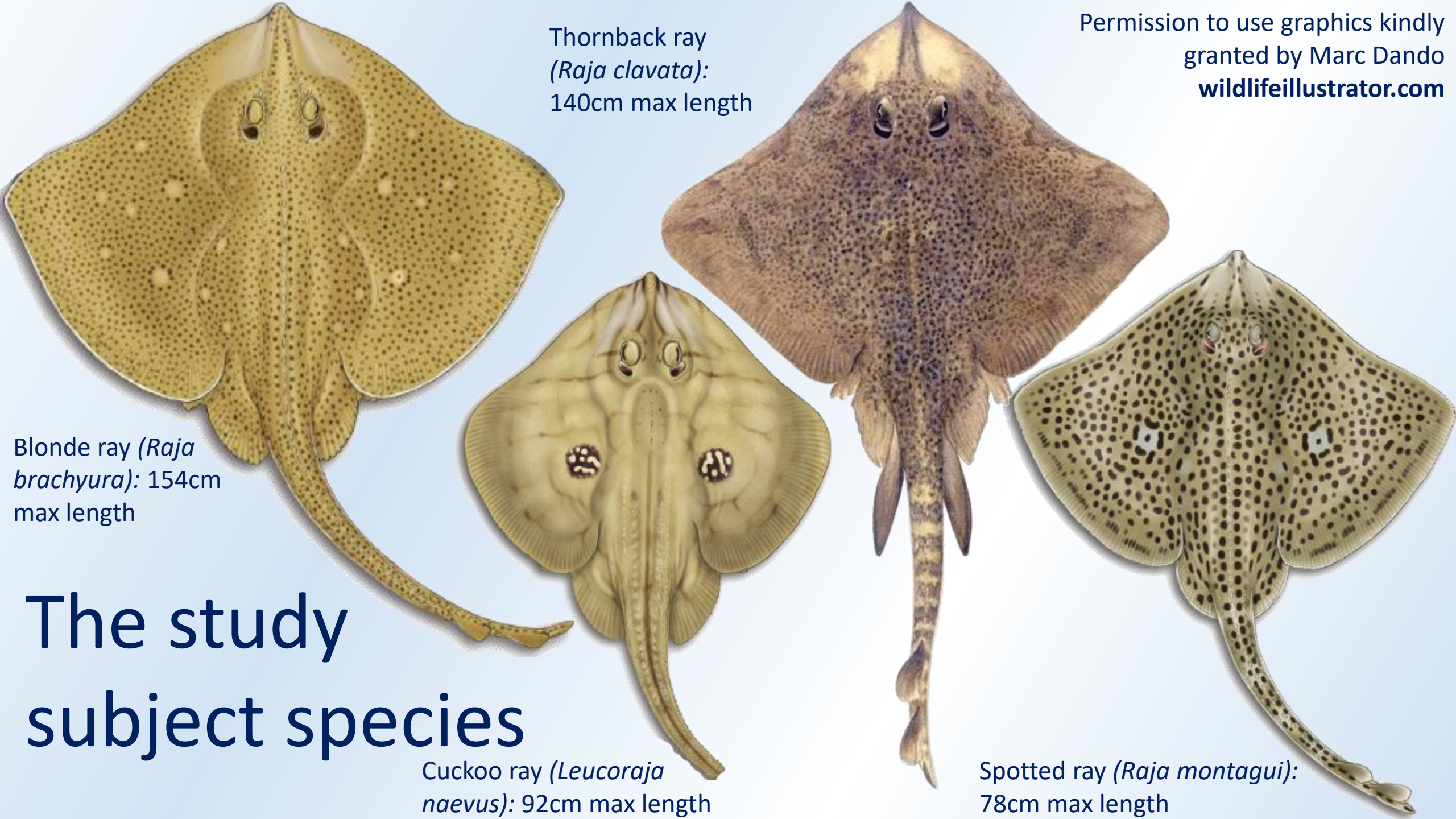
Maximum monthly bottom current speed 2010-2012: 22506 1185x1680m grids,MI



Distance from shore: map calculation



Grain size: ~250m minimum resolution, British Geological Survey (converted from sediment type classifications)



Permission to use graphics kindly granted by Marc Dando **wildlifeillustrator.com**

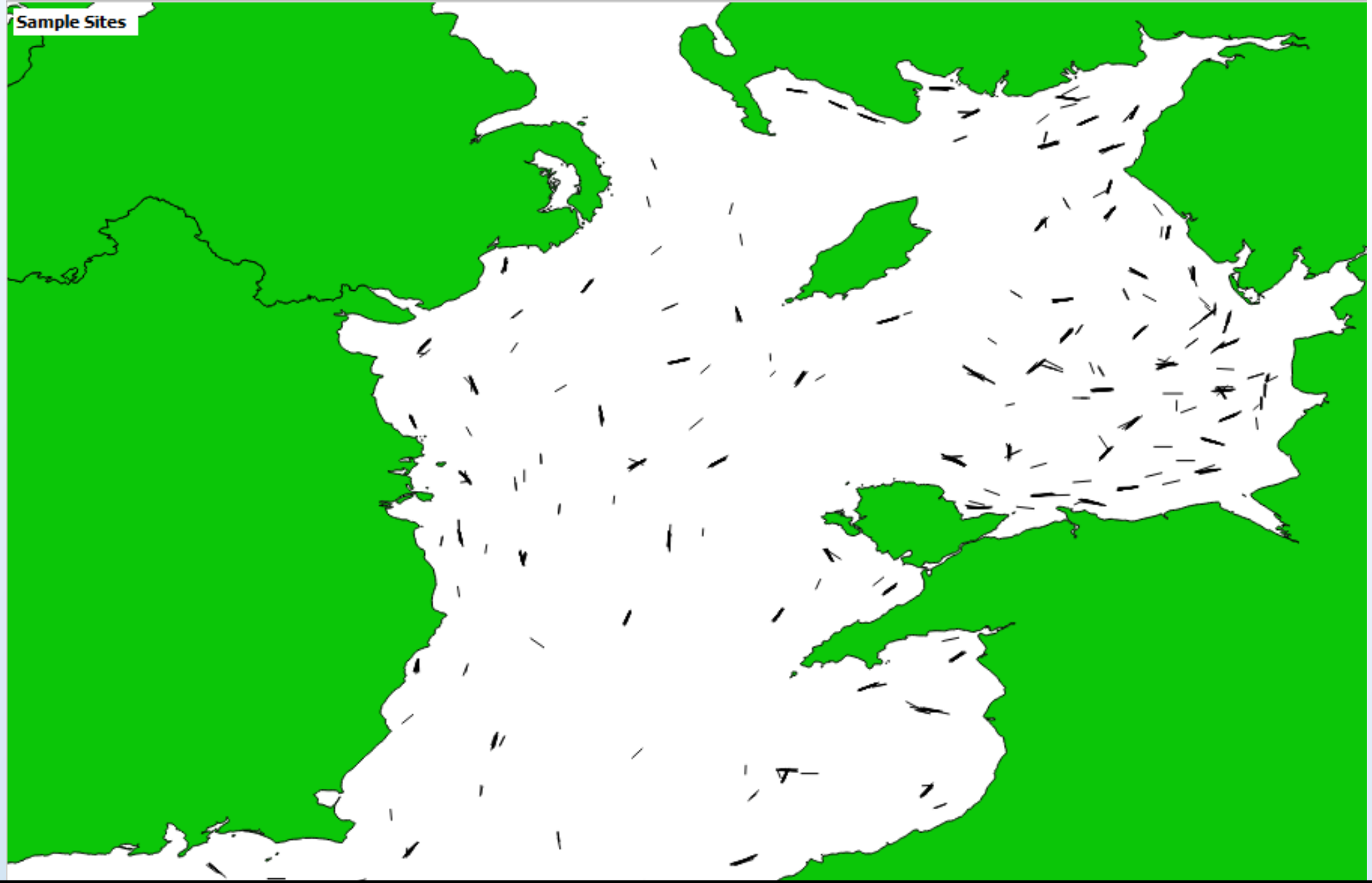
Thornback ray *(Raja clavata):* 140cm max length

Blonde ray *(Raja brachyura):* 154cm max length

The study subject species

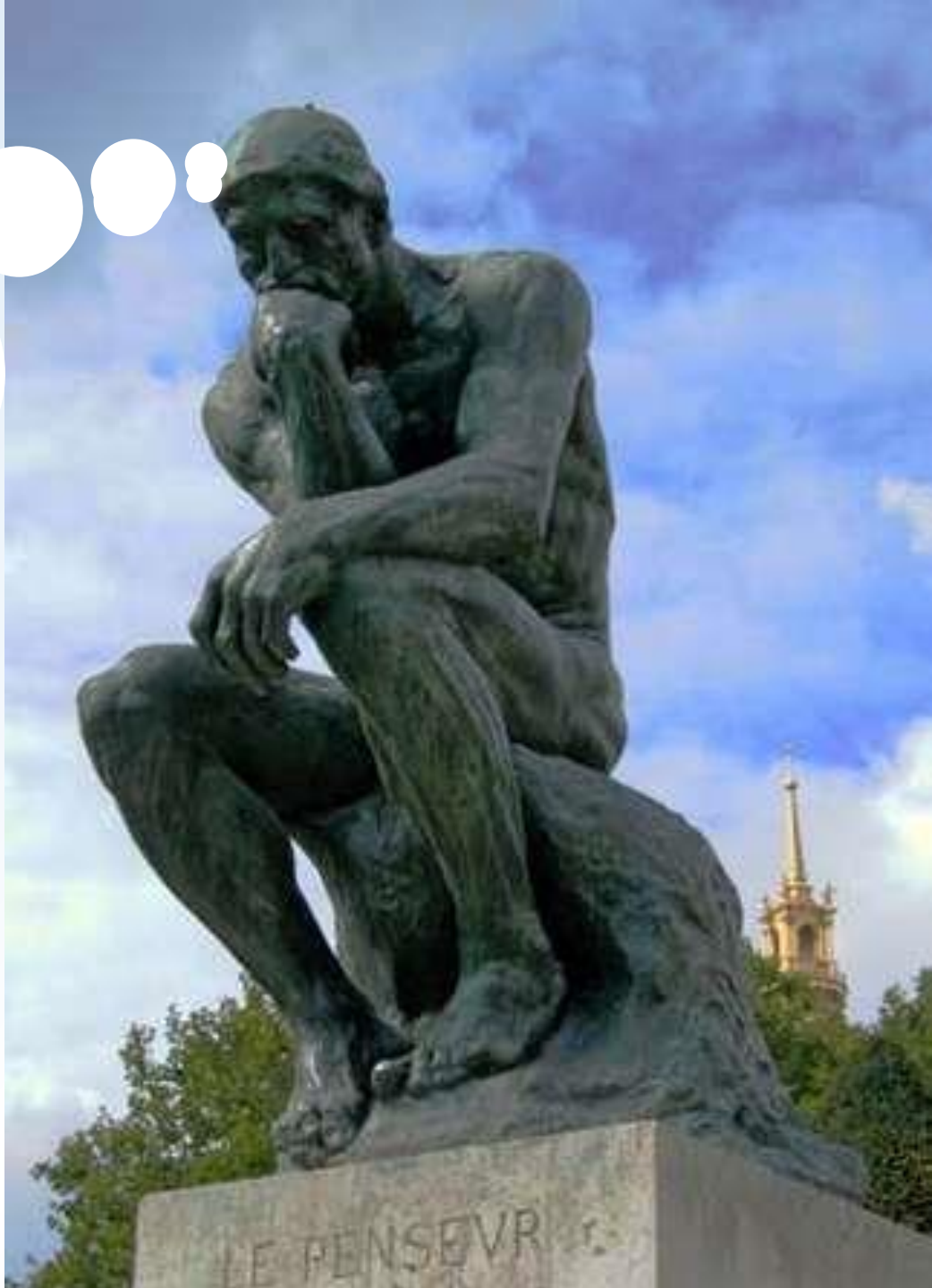
Cuckoo ray *(Leucoraja naevus):* 92cm max length

Spotted ray *(Raja montagui):* 78cm max length



Ray abundance at 1447 survey sites: ICES DATRAS, 1993-2012

At what point does making complex stats increasingly available risk a glut of experts testing and improving the models?



People in glass houses shouldn’t throw stones… maybe Elith et al would say the same about me!

ECOGRAPHY 29: 129–151, 2006

Novel methods improve prediction of species' distributions from occurrence data

Jane Elith*, Catherine H. Graham*, Robert P. Anderson, Miroslav Dudík, Simon Ferrier, Antoine Guisan, Robert J. Hijmans, Falk Huettmann, John R. Leathwick, Anthony Lehmann, Jin Li, Lucia G. Lohmann, Bette A. Loiselle, Glenn Manion, Craig Moritz, Miguel Nakamura, Yoshinori Nakazawa, Jacob McC. Overton, A. Townsend Peterson, Steven J. Phillips, Karen Richardson, Ricardo Scachetti-Pereira, Robert E. Schapire, Jorge Soberón, Stephen Williams, Mary S. Wisz and Niklaus E. Zimmermann

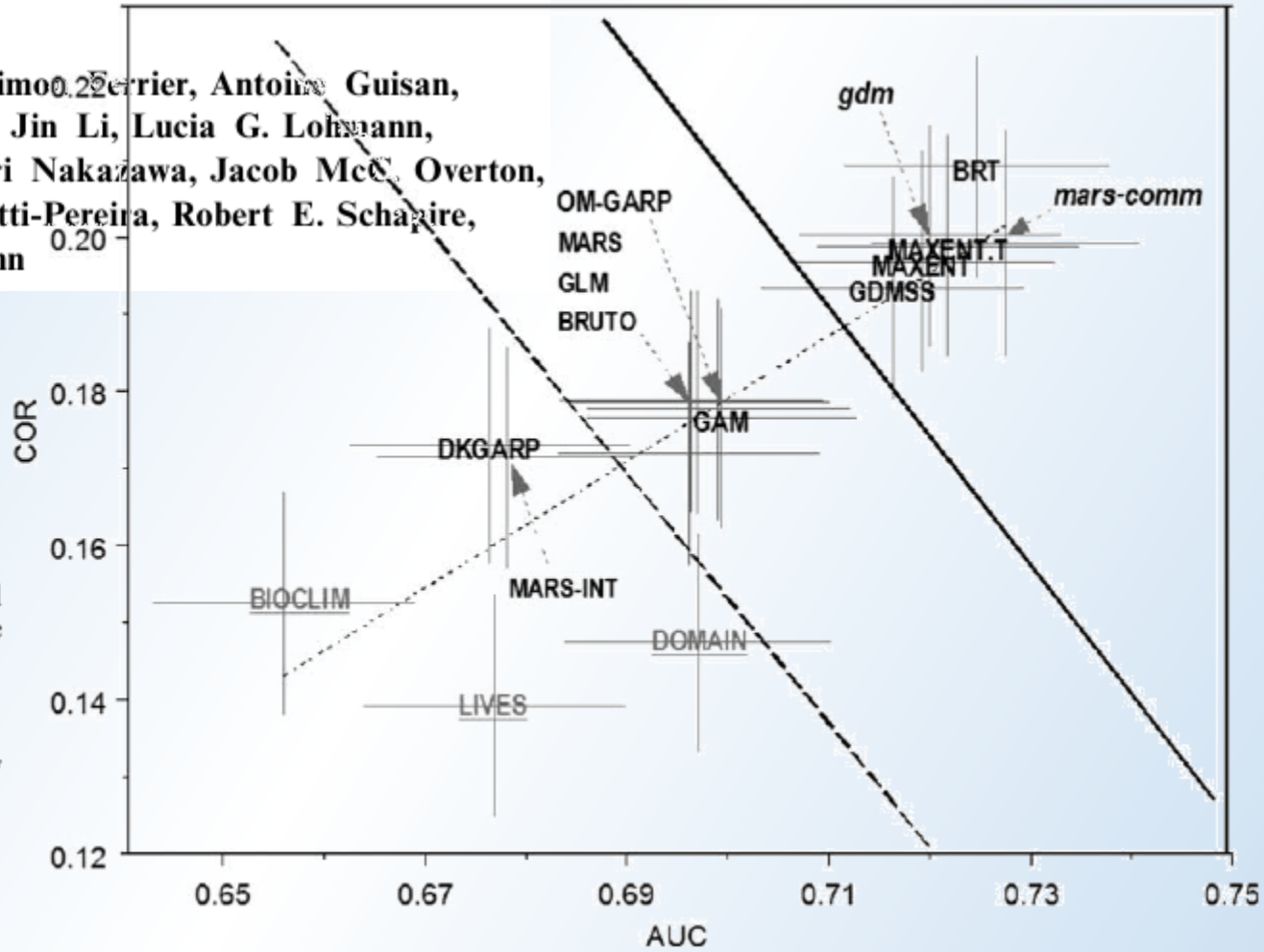


Fig. 3. Mean AUC vs mean correlation (COR) for modelling methods, summarised across all species. The grey bars are standard errors estimated in the GLMM (see Appendix), reflecting variation for an average species in an average region. The labels are broad classifications of the methods: grey underlined = only use presence data, black capitals = use presence and background samples, black lower case italics = community methods.

ToDo List: improvements, additions, bugs

- Parallelisation: the core BRT function is a sequential process i.e. single thread only, but could run both halves of a delta model simultaneously.
- OS compatibility
- Swept area AND Spatial error implicit in input data
- Processing time estimate
- Parameter optimisation

Brush Size:

Brush Shape:

Circle | Square | Drag Selection

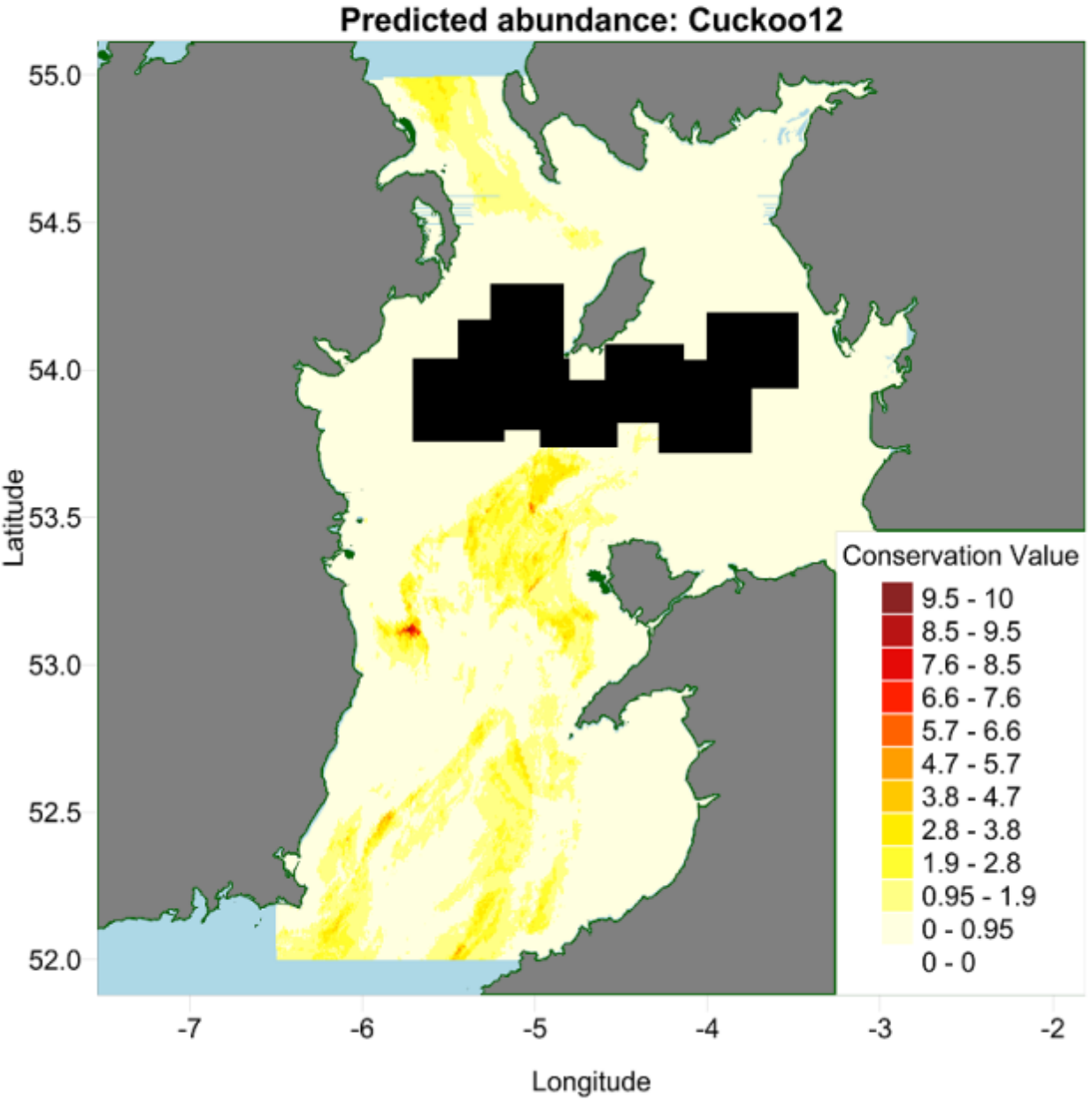
Closed area

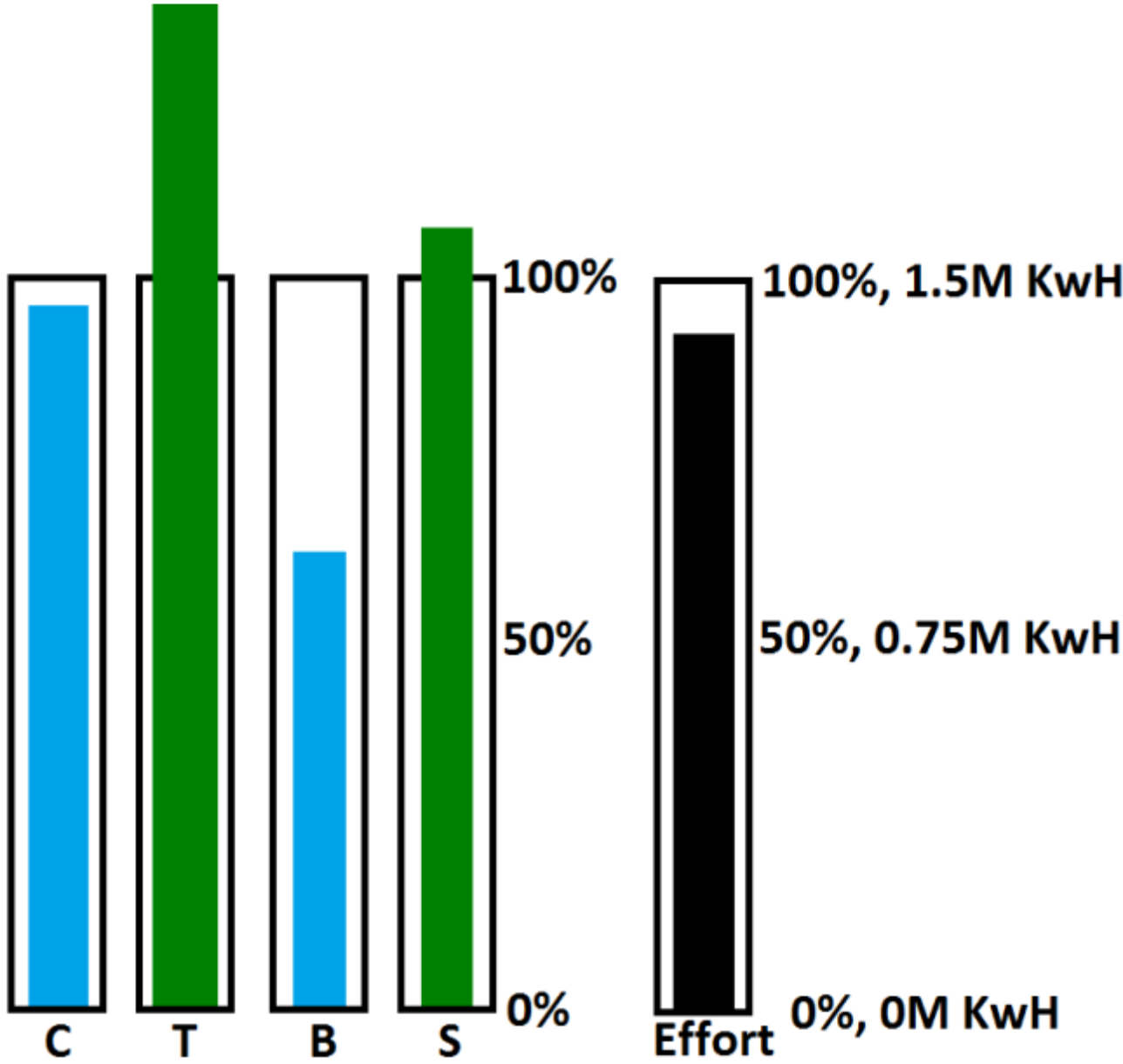
open area

Save/export

Open/import

Paper 5 concept graphics





What I'm doing next

- Farallon Institute, Petaluma, CA
- Developing a population dynamics model on forage fish (central northern stock of northern anchovy) abundance in relation to environmental conditions , fisheries exploitation & trophic (predator-prey) interactions in the Southern California Current System using available acoustic & trawl survey data (CalCOFI)
- Explain state shifts
- Non-stationary model, Bayesian TMB? Spatial? Range expansion / contraction
- Sardine eat anchovy eggs…

Thanks. Any questions?

- Entire project coded in R & requires minimal R knowledge **github.com/SimonDedman/gbm.auto**
- Code / figures / contact / everything: **simondedman.com** **simondedman@gmail.com**
- Ecological Modelling 312 (2015) 77–90: Modelling abundance hotspots for data-poor Irish Sea rays
- Fishes 2 12 (2017)1–22: Advanced spatial modelling to inform management of data-poor juvenile & adult female rays
- ICES Journal of Marine Science 74:2 (2017) 576-587: Towards a flexible Decision Support Tool for MSY-based Marine Protected Area design for skates and rays
- PLoS ONE 12(12): e0188955: Gbm.auto: a software tool to simplify spatial modelling and Marine Protected Area planning
- Bangley *et al.*: PLoS ONE (in Review): Delineation and Mapping of Coastal Shark Habitat within Pamlico Sound, NC
- Burke *et al.*: In Prep: Spatial analysis review of BRUVs data
- Grimmel *et al.*: In Prep: Assessment of Faunal Communities and Ecosystem Interactions within a Shallow-water Lagoon using BRUVs
- Please let me know criticisms/praise/suggestions by email or in person. Thanks!

Thanks to Dr Chuck Bangley, beta tester

Permission to use ray graphics kindly granted by Marc Dando **wildlifeillustrator.com** . All maps by SD

GMIT INSTITIÚID TEICNEOLAÍOCHTA NA GAILLIMHE-MAIGH EO GALWAY-MAYO INSTITUTE OF TECHNOLOGY

Marine Institute Foras na Mara

myfish Maximising yield of fisheries while balancing ecosystem, economic and social concerns

SEVENTH FRAMEWORK PROGRAMME

EUROPEAN REGIONAL DEVELOPMENT FUND