Data-driven bioregions for local ecosystem context in species distribution models

Andrew McMillan, Simon Fraser University
Anders Knudby, University of Ottawa

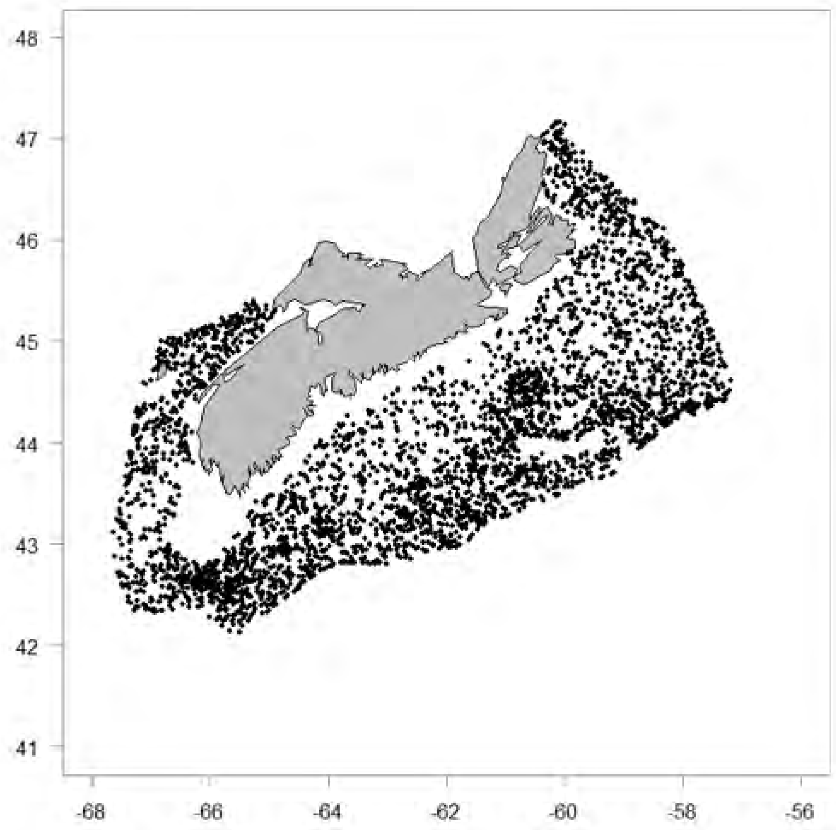
- Species distribution modeling is based on empirical relationships observed between species presence/absence and environmental variables

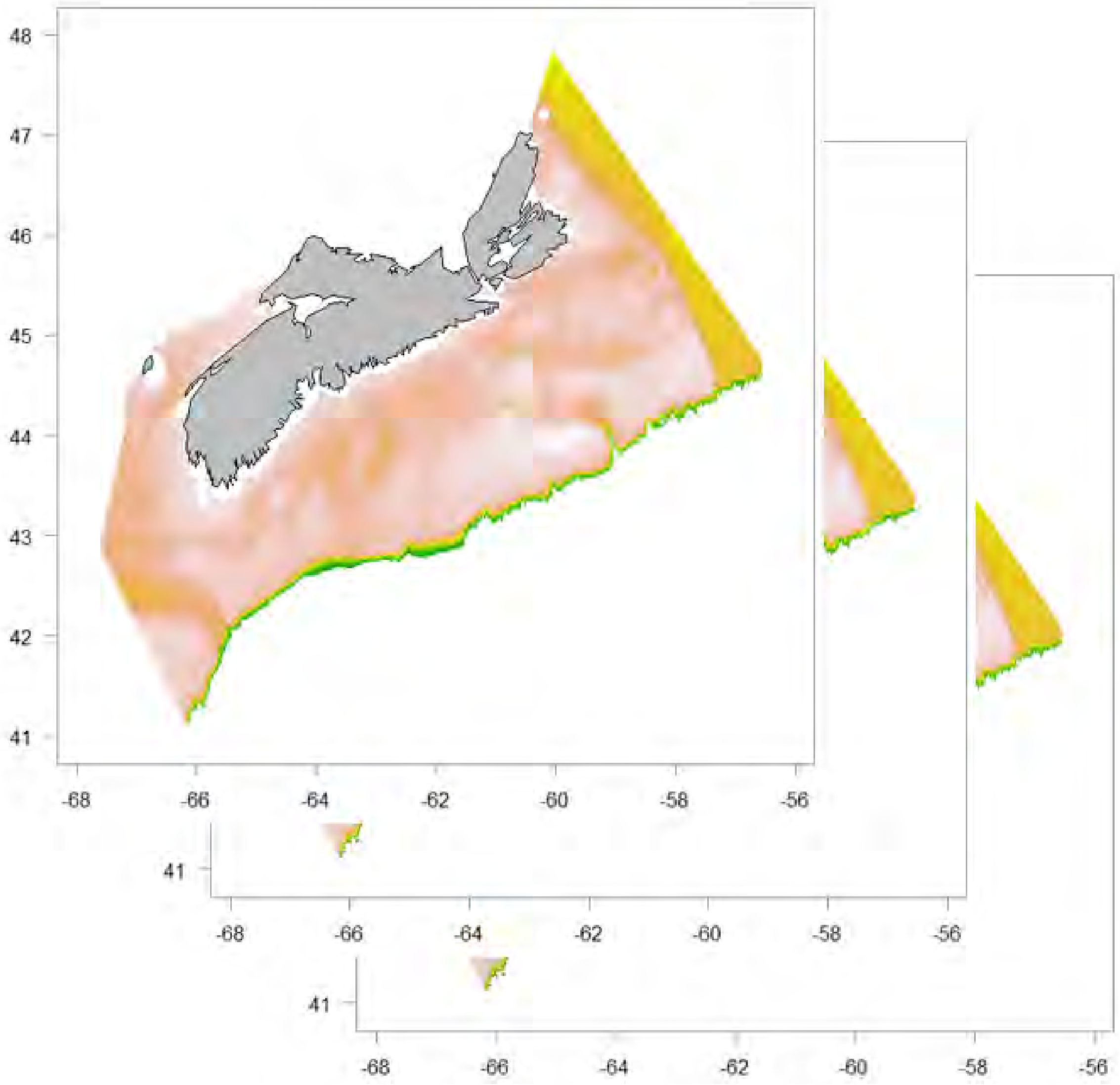
- These relationships change across space because they may:
    - be modified by interaction effects (e.g. water temperature may regulate the effect aragonite/calcite saturation has on species presence/absence)
    - act as proxies for unmeasured environmental variables (e.g. depth may act as a proxy for light availability, but the exact relationship depends on turbidity)
    - be influenced by the local species assemblage

- Regional calibration of SDMs may therefore improve their appropriateness

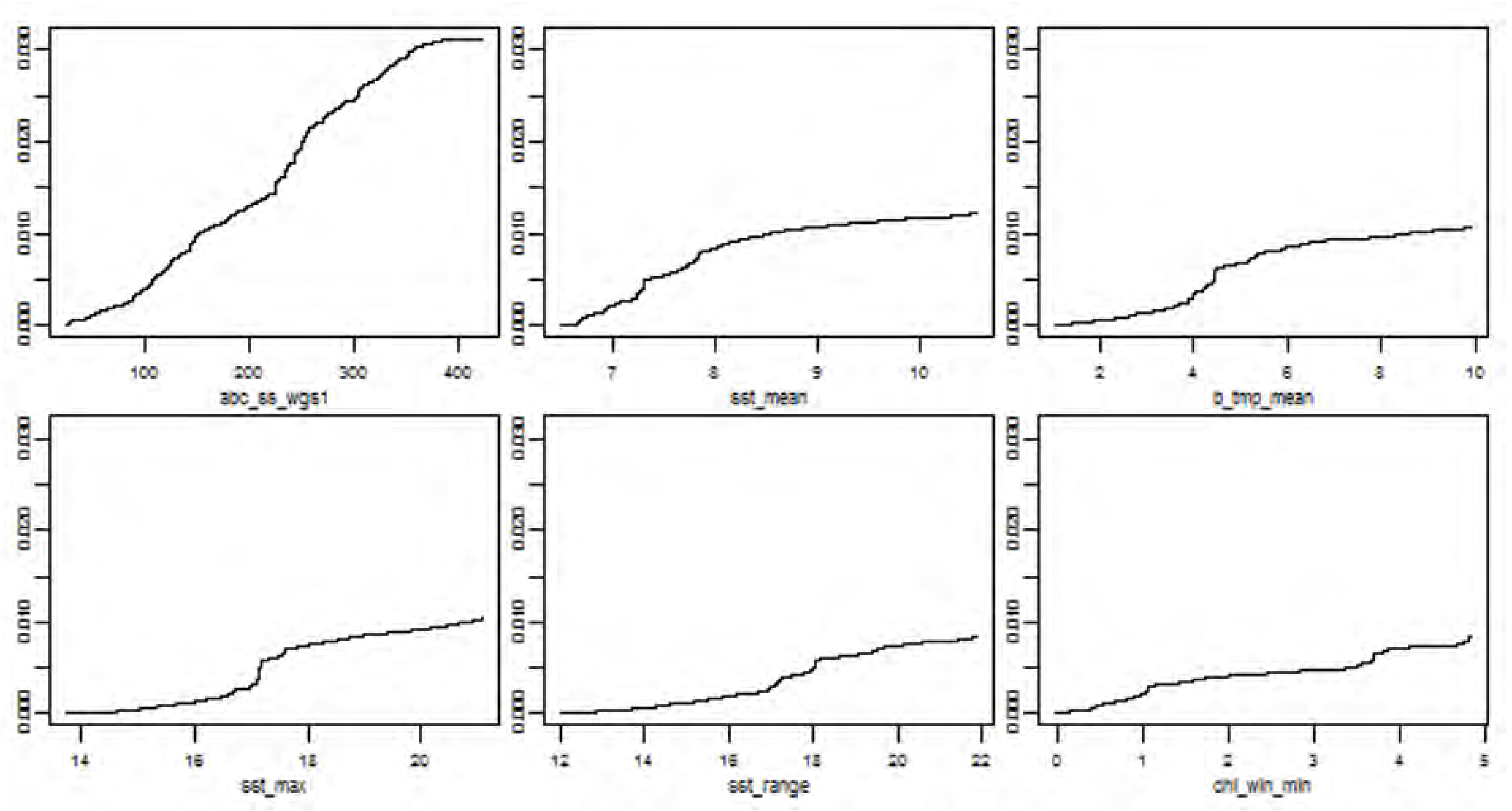
- One question, then, is how to define the regions?

Step 1: Gather data layers

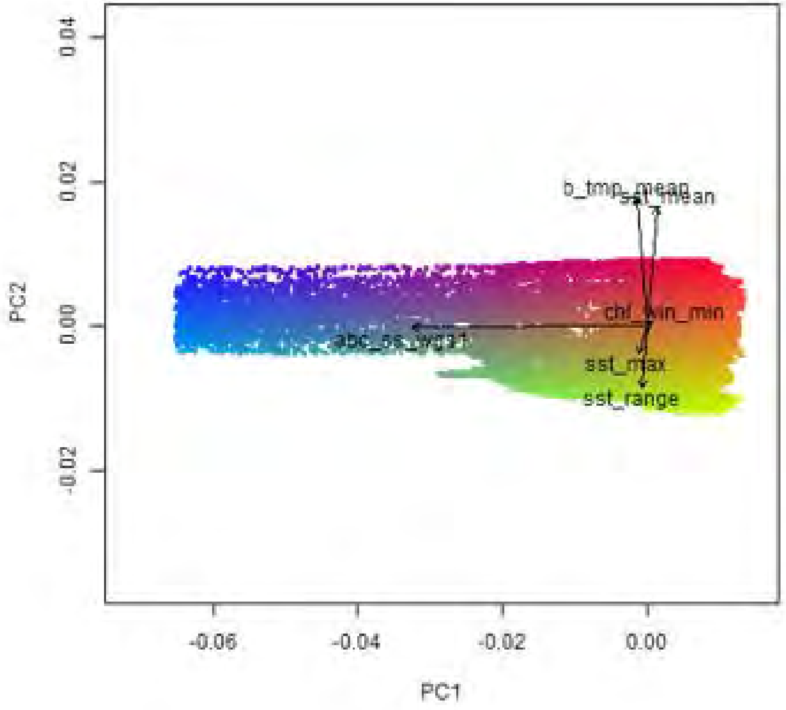


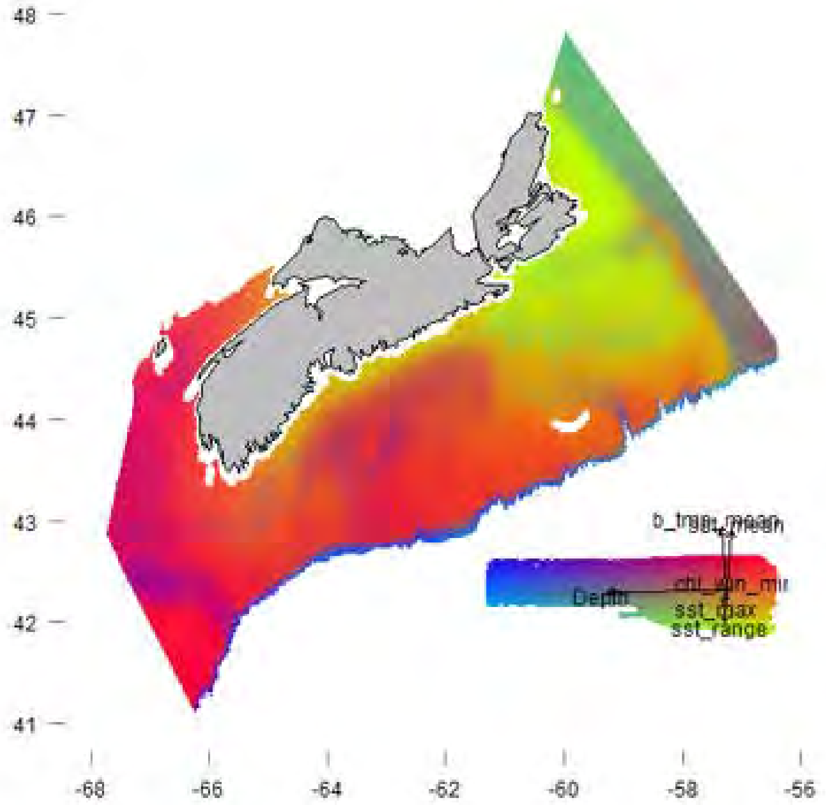


Step 2: Calibrate Random Forest and Gradient Forest models, calculating average of species response functions for each environmental variable (i.e. compositional turnover along environmental gradients)

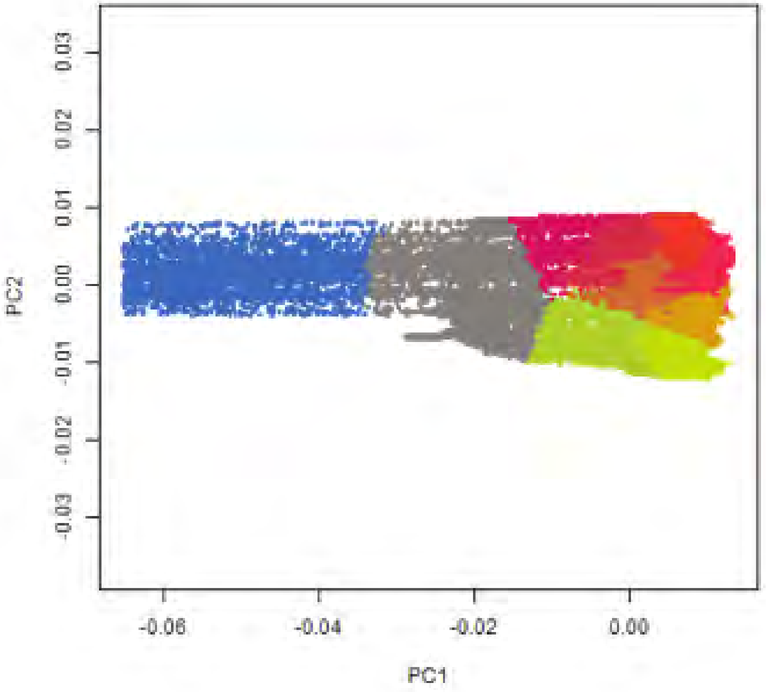


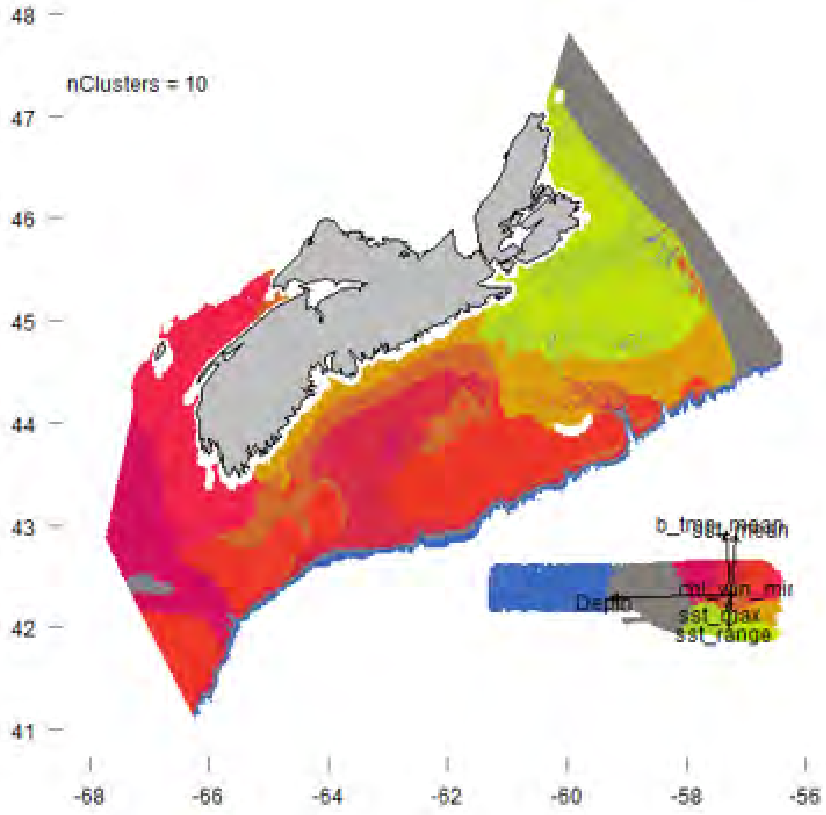
Step 3: Run PCA on predicted species turnover, and use principal components to map predicted species assemblage groups





Step 4: Run clustering algorithm on PCA scores and apply clusters to map





Results:

- Depth and bottom temperature have the greatest influence on species turnover
- Regionalization improves predictions (AUC) for some species, worsens predictions for others
- Is better than alternative regionalization schemes based on depth, latitude, longitude or geographic clustering for some species, but worse for other species

Advantages:

- Method entirely data driven
- Works well to regionalize by species assemblage

Challenges:

- Method ignores actual geography, (e.g. current systems)
- Number of clusters determined subjectively
- Trade-off between regionalization, number of clusters, and number of calibration data in each cluster
- Relies on comprehensive and standardized data set, typical only from trawl surveys