



Methods and Tools for Spatial Modeling

Elliott Hazen, Ben Best, Jason Roberts, Patrick Halpin

MATE / Satellite class

August 8, 2010



NICHOLAS SCHOOL OF THE
ENVIRONMENT AND EARTH SCIENCES
DUKE UNIVERSITY



Our Focus / Biases

- Disciplinary
 - Space (and Time)
 - Environment + Prey
 - Statistics & prediction

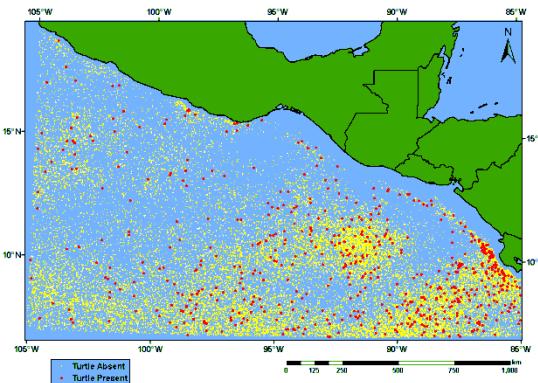
My Focus

- Disciplinary
 - Space (and Time)
 - Environment + Prey
 - Statistics to prediction
- Flowchart



Modeling habitat (overview)

Distribution data, e.g. presence/absence



Presence only data, e.g.

- Vessels of opportunity
- Hydrophones

Presence / absence, density data

- Survey sightings w/ effort
- Bycatch

Event based data

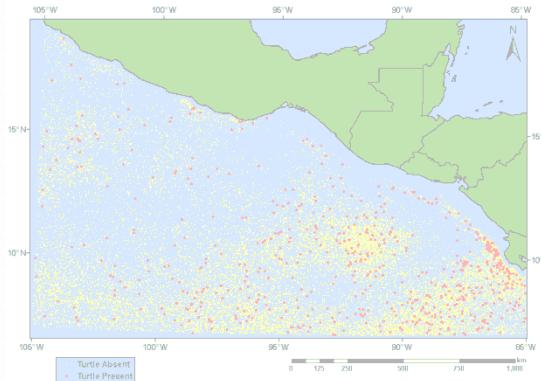
- Focal follows
- Short term tag data

Movement data

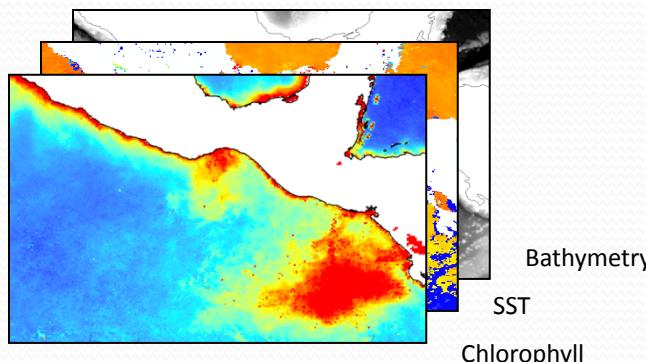
- Short and long term tag data

Modeling habitat (overview)

Distribution data, e.g. presence/absence



Sampled predictive data



In situ data

- Continuous data – surface sensors, fisheries acoustics, ADCP
- Station data – CTDs, trawls

Remotely sensed data

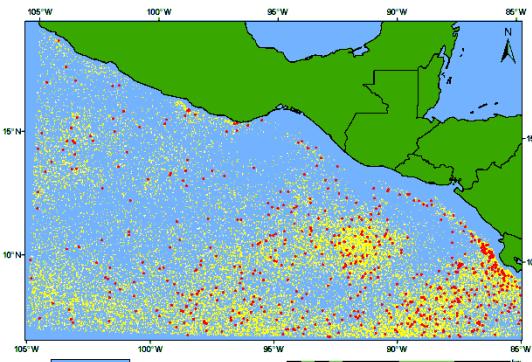
- SST, SSH, Chl - Data centers

Physical / spatial data

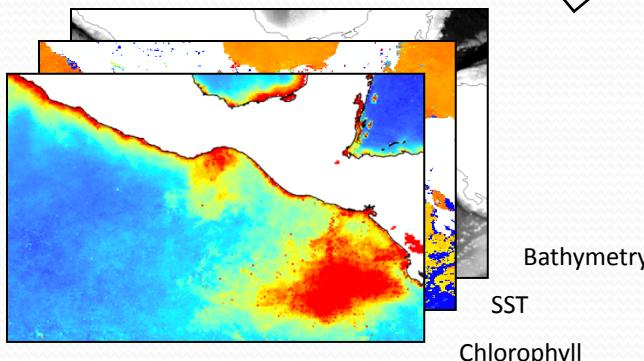
- Bathymetry, distance from feature (e.g. slope, shore, break, front)

Modeling habitat (overview)

Distribution data, e.g. presence/absence



Sampled predictive data

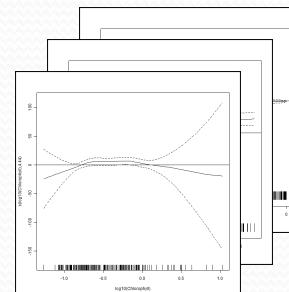


Statistical models

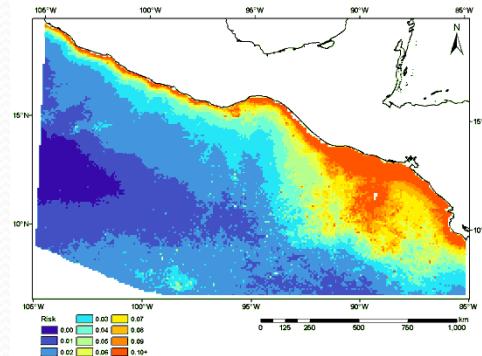
$$g(\mu) = \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m$$

Fit

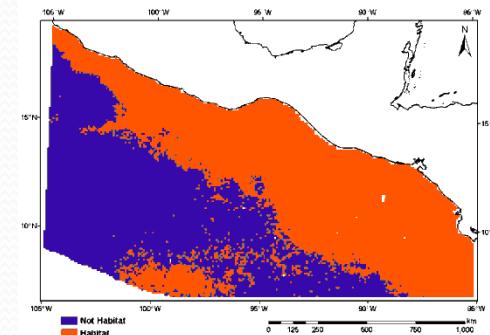
Predict



Probability of occurrence predicted from environmental covariates

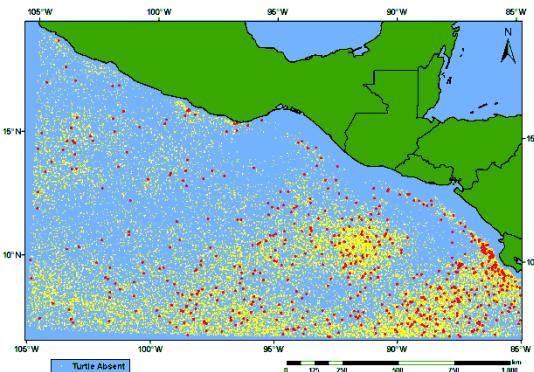


Binary classification

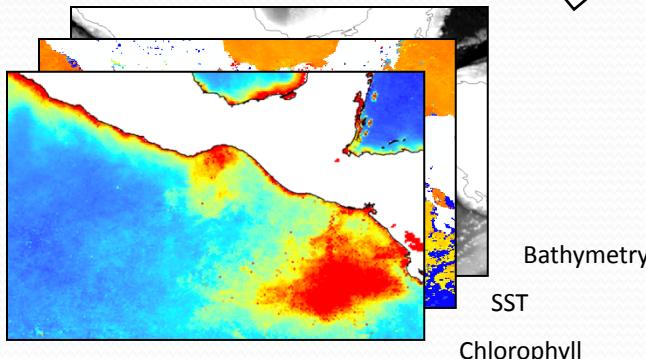


Model selection

Distribution data, e.g. presence/absence

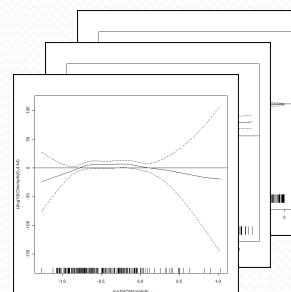


Sampled predictive data



Statistical models

$$g(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$



Akaike's
Information
Criterion

$$\longrightarrow \text{presence} \sim s(\text{SST}) + s(\text{prey.density})$$

Bayesian
Information
Criterion

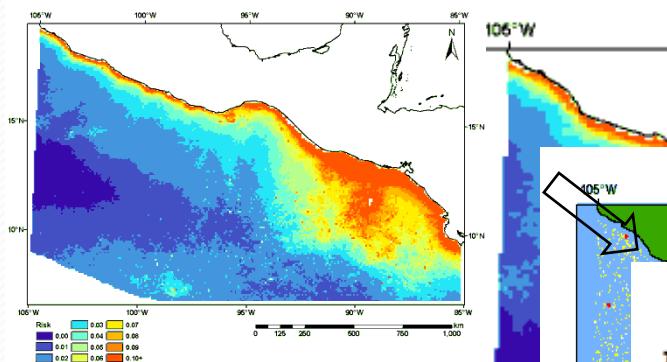
“Final” model

$$\begin{aligned} AIC &= 2k - 2\ln(L) \\ BIC &= 2\ln(L) + k\ln(n) \end{aligned}$$

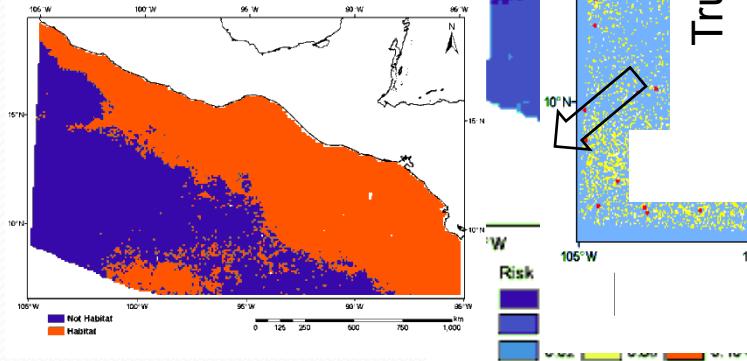
n – sample size
 k – number of parameters
 L – Likelihood function

Evaluating habitat models

Probability of occurrence

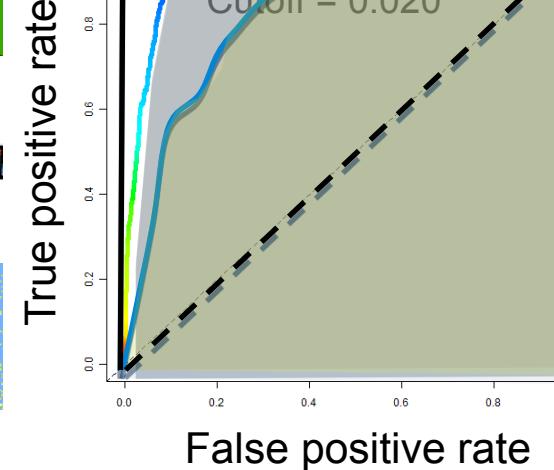


Binary classification

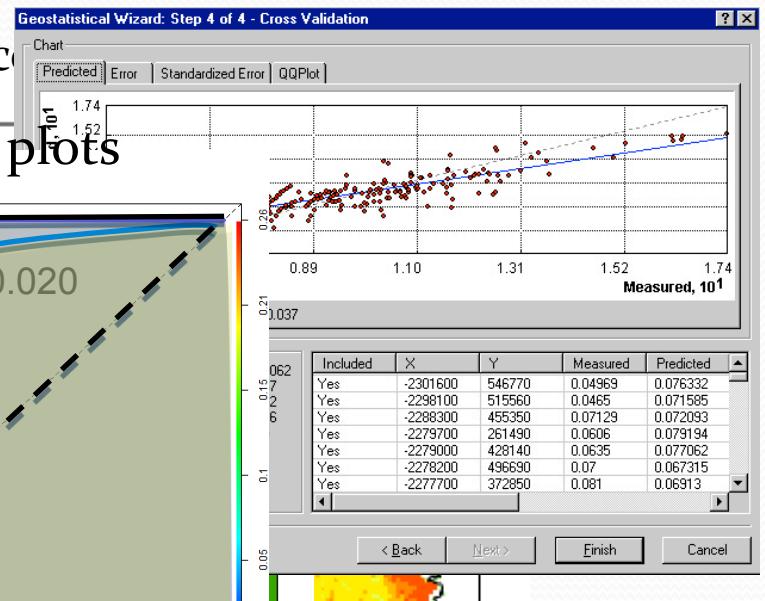


Probability of occurrence

ROC plots



Cross-validation

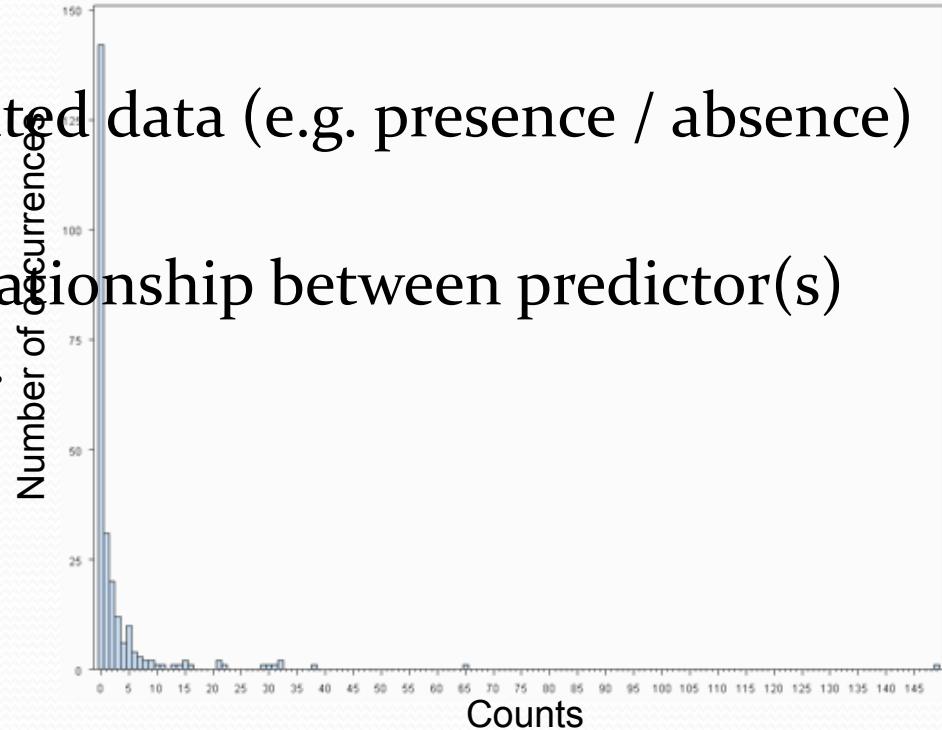


Types of statistical models

- There are many, and constantly changing / growing
- Correlation/Regression techniques – GLMs, GAMs (Austin 2002), Mixed models (Wood 2006), regression trees & random forests (Breiman 2001)
- Ordination – Multivariate dimensional scaling, e.g. CCAs (Guisan et al. 1999),
- Maximum Entropy models – species distributions “closest to uniform” (Phillips et al. 2006)
- Recent reviews of modeling approaches (Redfern et al. 2006, Elith et al. 2006, Dormann et al. 2007, Aarts et al. 2008)

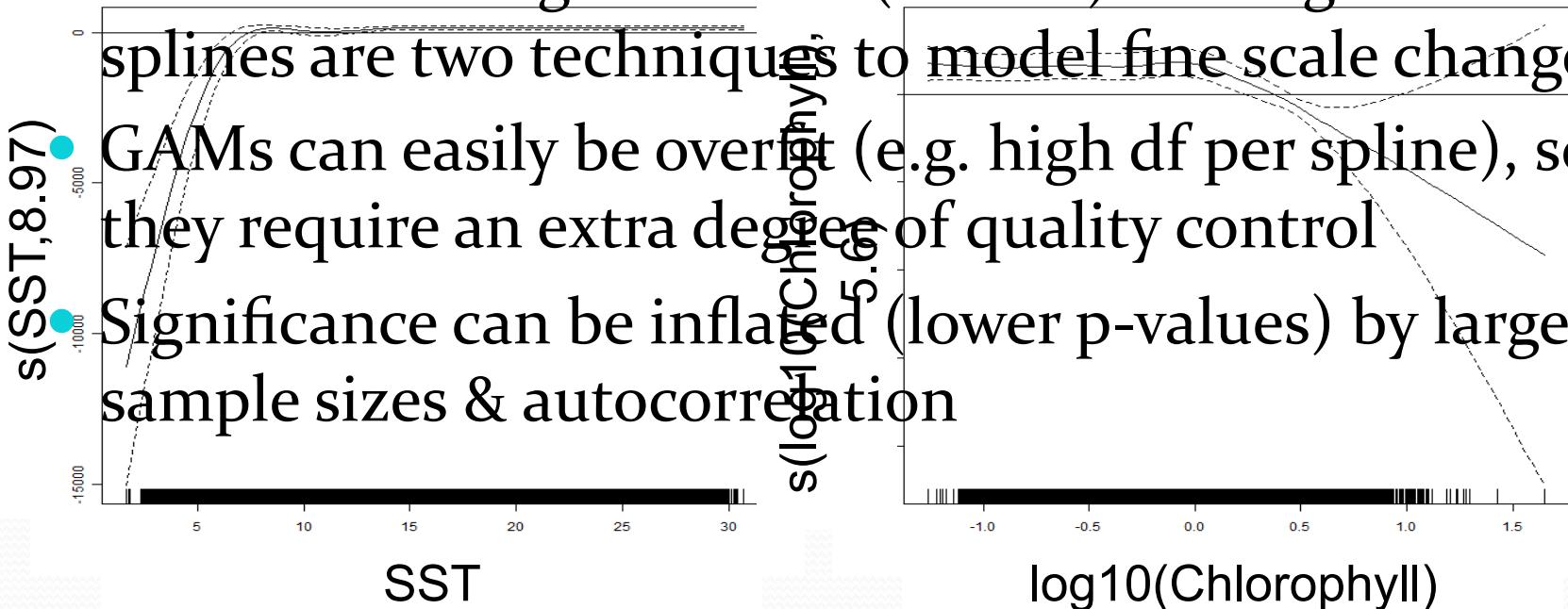
Generalized Linear Models

- GLMs are an extension of linear models, $y \sim f(x_1, x_2, \dots, x_n) + \varepsilon$ using MLE and a link function
 - For data with many zeroes (e.g. count data), Poisson – log link
 - For binomially distributed data (e.g. presence / absence) - logit link function
 - Relies upon a linear relationship between predictor(s) and response variables.



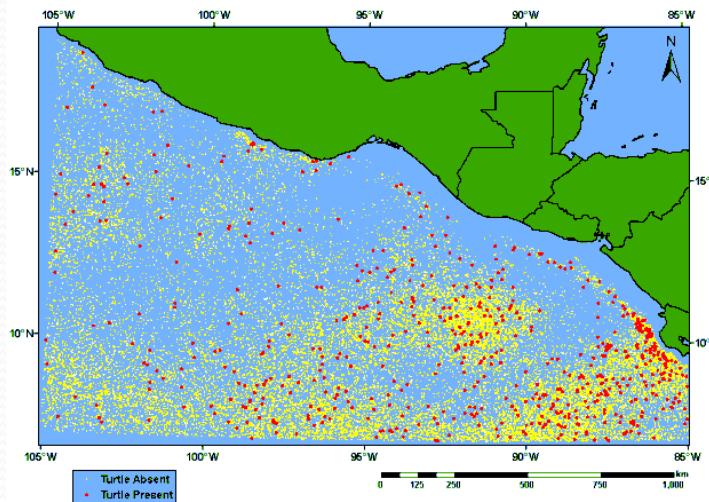
Generalized Additive Models

- GAMs can use a combination of parametric and non-parametric functions ($y \sim A + f(x_1) + f(x_2) + \dots + f(x_n) + \varepsilon$)
- Local smoothing functions (LOESS) and regression splines are two techniques to model fine scale changes
- GAMs can easily be overfit (e.g. high df per spline), so they require an extra degree of quality control
- Significance can be inflated (lower p-values) by large sample sizes & autocorrelation



Spatial autocorrelation

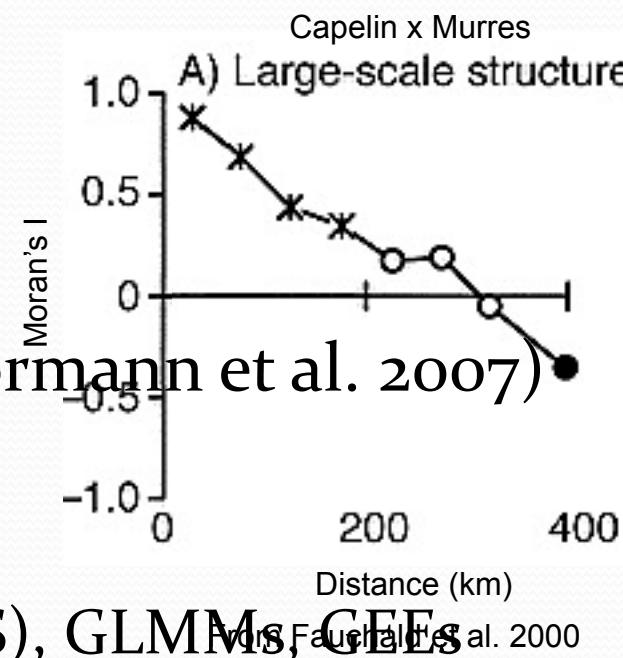
- Why do we care?
- Model assumption is independence of data points
- Spatial autocorrelation may bias model results (see Segurado et al. 2006)



Albatross track leaving French Frigate Shoals
From OBIS-SEAMAP (Shaffer et al. 2005)

Spatial autocorrelation

- Ways to test for it
 - Geary's C (0 to 2)
 - Moran's I (-1 to 1)
- Many methods to model it (Dormann et al. 2007)
 - Autocovariate regression & spatial eigenvectors
 - Generalized least squares (GLS), GLMMs, GEEs
 - Partial Mantel's tests (Legendre and Legendre 1998)



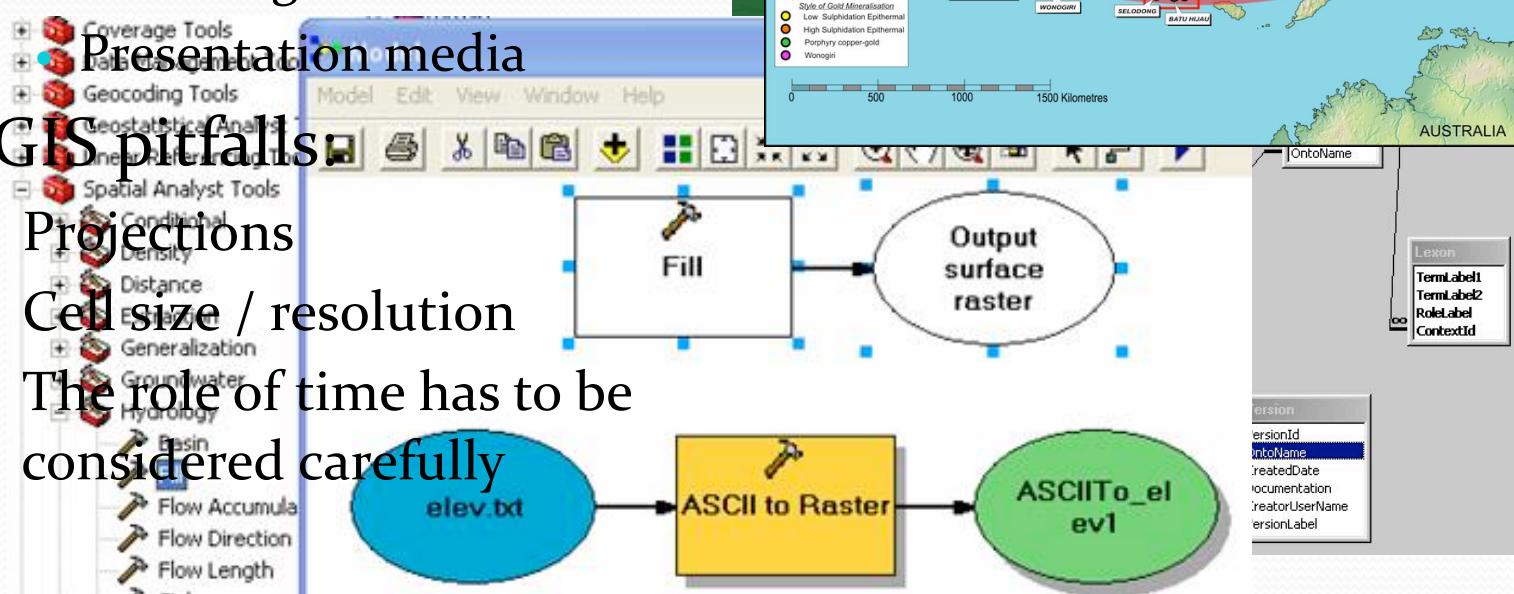
Dormann et al. 2000

Useful software packages

- MATLAB / IDL – multipurpose scientific programming language; PERL
 - WinBugs – toolset for bayesian analysis
 - EcoPath / EcoSim – Mass balance models
 - R / S⁺ / SAS – statistical programming language
 - Python – scripting language used by Arc
 - ArcGIS Desktop – Geographic Information System
 - Model builder
 - Hawth's tools, Biomapper, MGET toolbox, EDC
-

Introduction to GIS

- What is a GIS used for?
 - Spatial analyses
 - Relational databases
 - Modeling
- ArcGIS pitfalls:
 - Projections
 - Cell size / resolution
 - The role of time has to be considered carefully



What is EDC?



- Allows users to connect to data servers from within ArcGIS and download environmental data
- Works with THREDDS/OPeNDAP servers to provide feature and raster data
- Incorporates time in ArcGIS (great for videos)



EDC installation

- <http://www.pfeg.noaa.gov/products/EDC/index.html>
- Requires:
 - ArcGIS 9.2 sp 3 or higher (not yet 10)
 - Java Run-Time Environment 6
 - .NET Framework 2.0 or 3.0
 - .NET Support for ArcGIS Desktop



EDC workflow

- Finding data

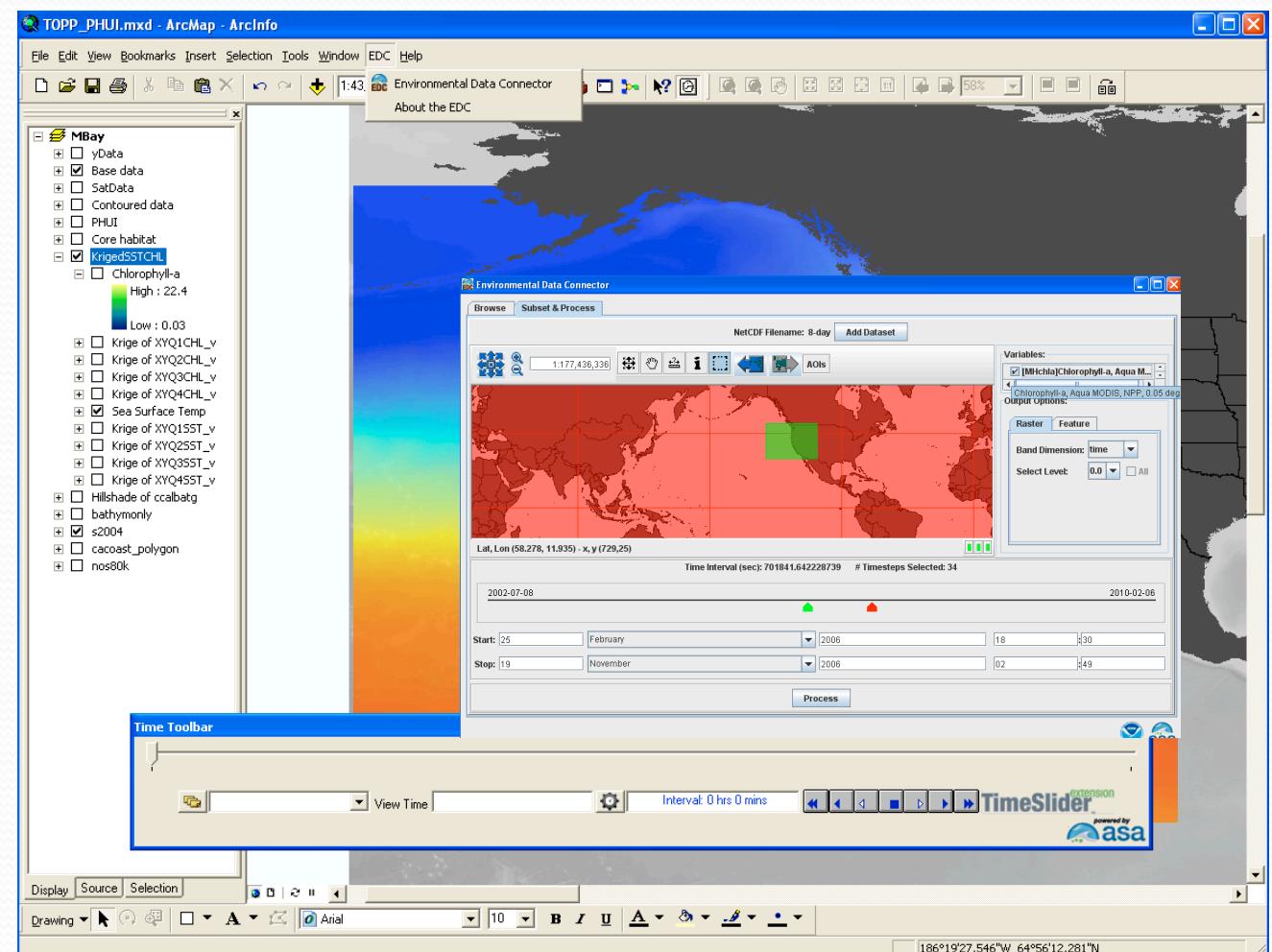
The screenshot shows the ArcMap application interface with several windows open:

- TOPP_PHUI.mxd - ArcMap - ArcInfo**: The main map window displays a coastal area with various data layers.
- MBay**: A legend and data source list for the MBay layer, including "KrigedSSTCHL" which is checked.
- Environmental Data Connector**: A floating window with the title "About the EDC".
- Environmental Data Connector**: A separate window showing the "Direct Access URL" tab with a list of URLs for different datasets like MODIS, NPP, and COASTWATCH.
- Time Toolbar**: A toolbar at the bottom for managing time intervals, with "TimeSlider" branding.



EDC workflow

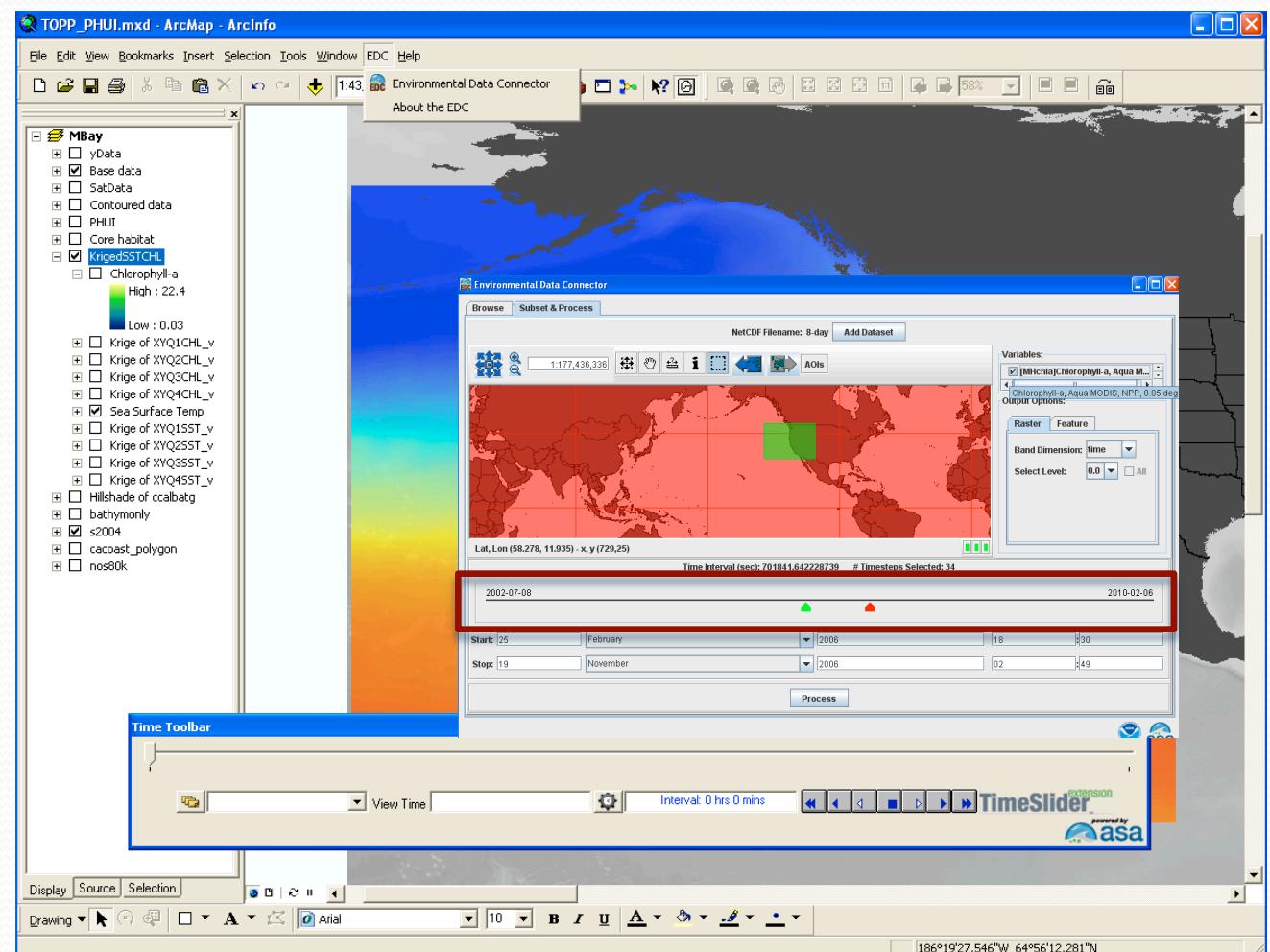
- Finding data
- Subset data





EDC workflow

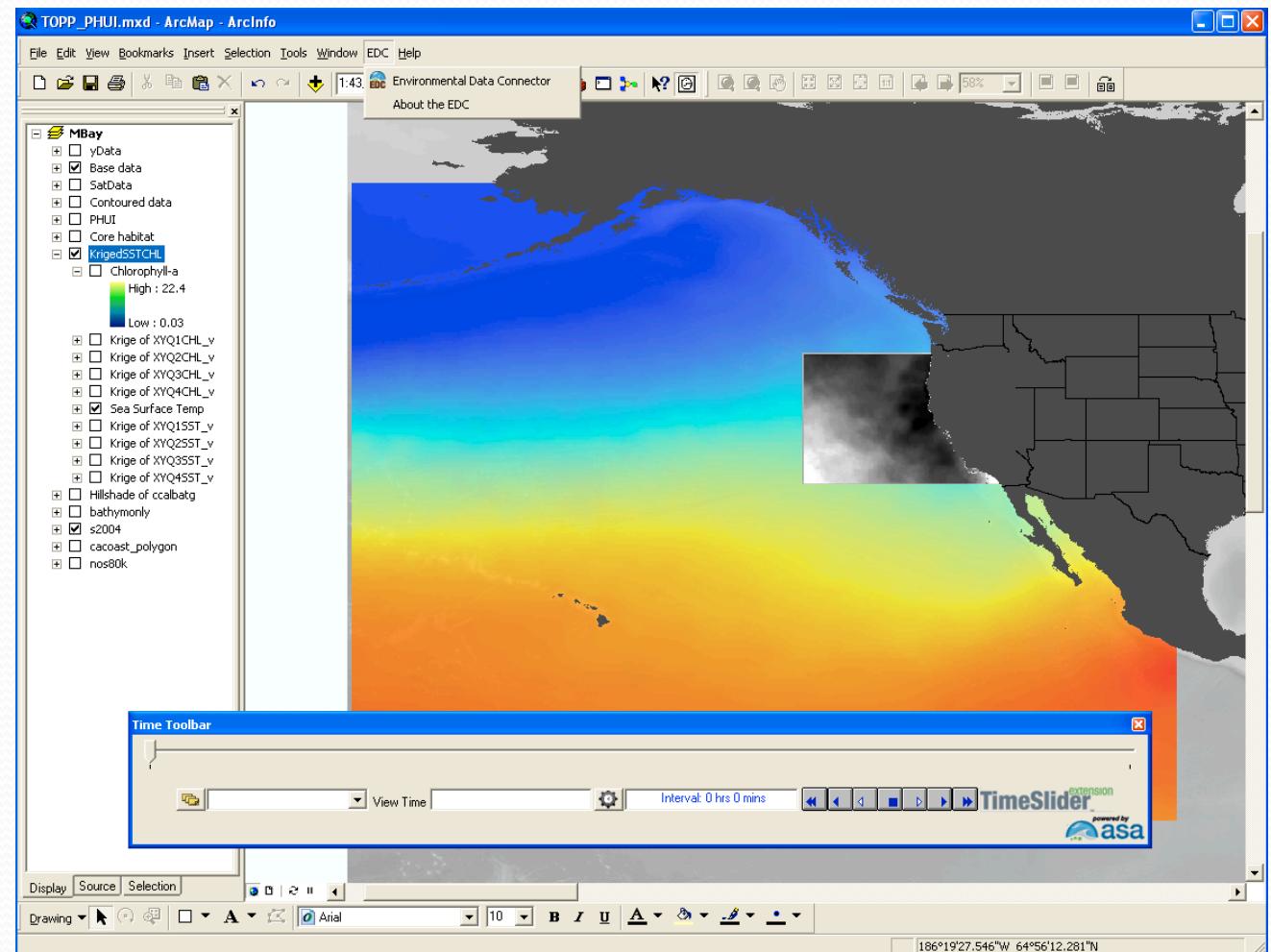
- Finding data
- Subset data
- Choose time





EDC workflow

- Finding data
- Subset data
- Choose time
- Process



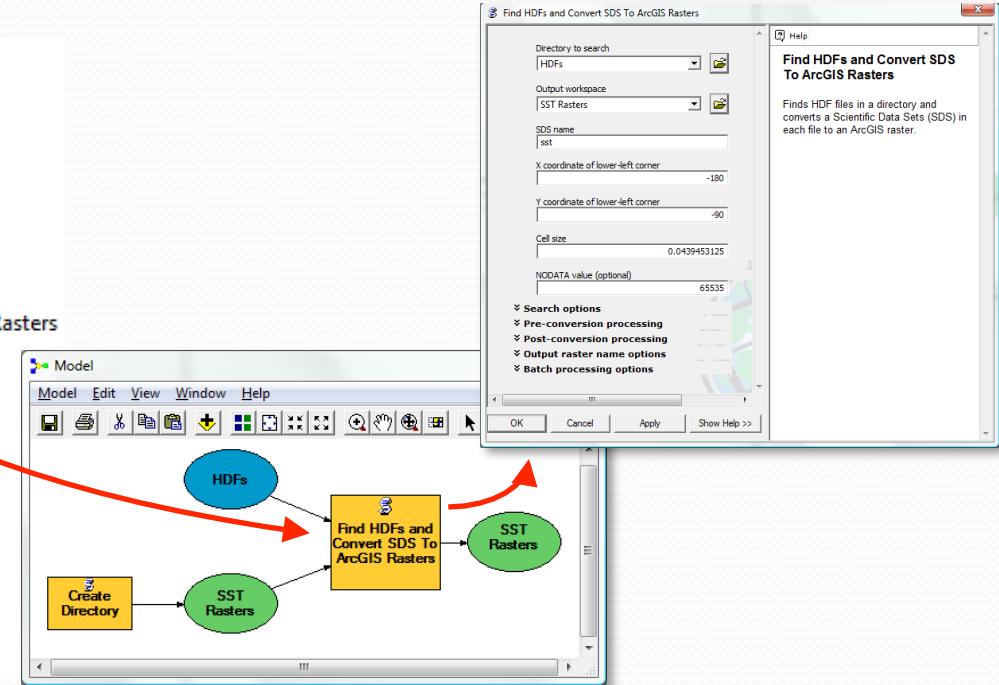
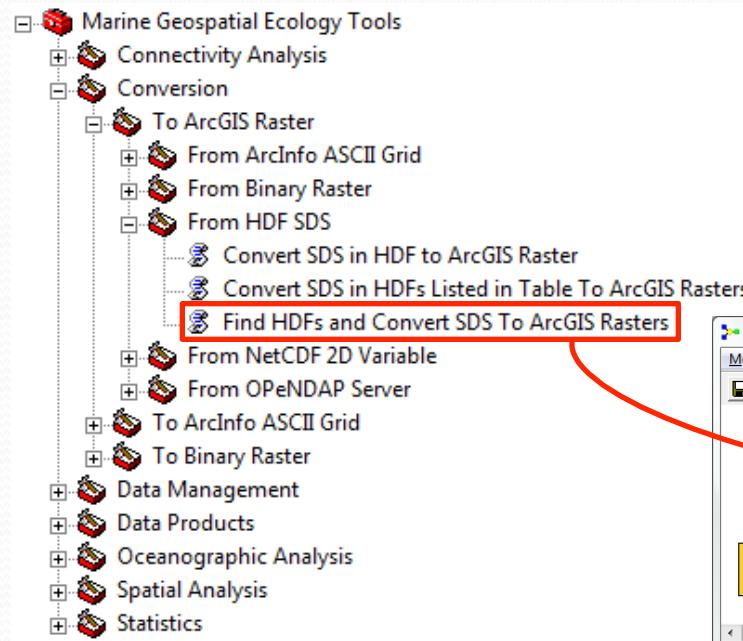
What is MGET?

<http://code.env.duke.edu/projects/mget>

- A collection of geoprocessing tools for marine ecology
 - Oceanographic data management and analysis
 - Habitat modeling, connectivity modeling, statistics
 - Highly modular; designed to be used in many scenarios
 - Emphasis on batch processing and interoperability
- Free, open source software
- Written in Python, R, MATLAB, C#, and C++
- Minimum requirements: Win XP, Python 2.4
- ArcGIS 9.1 or later currently needed for many tools
- ArcGIS and Windows are only non-free requirements

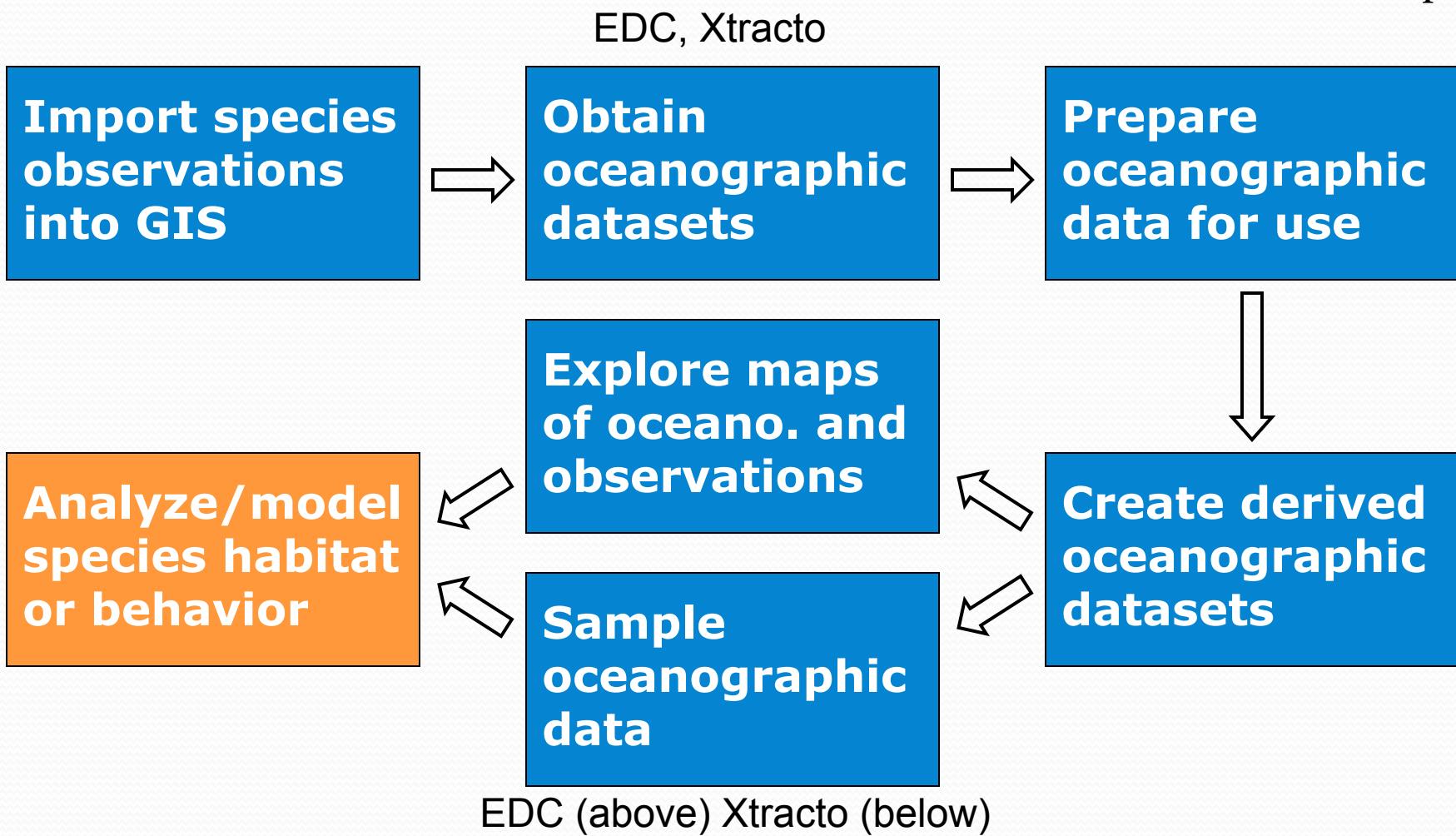
MGET interface in ArcGIS

- Expand the toolbox to find the tools
- Double-click tools to execute directly, or drag to geoprocessing models to create a workflow

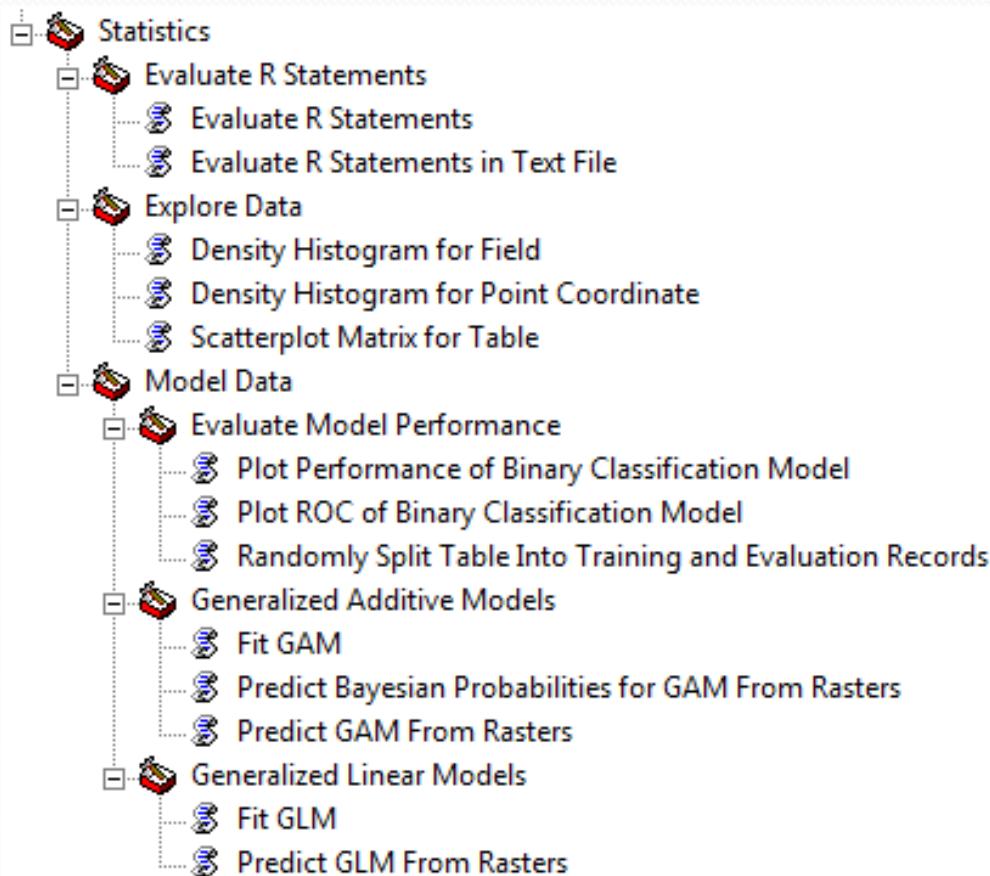


Simplified workflow

MGET includes tools that assist with all of these steps



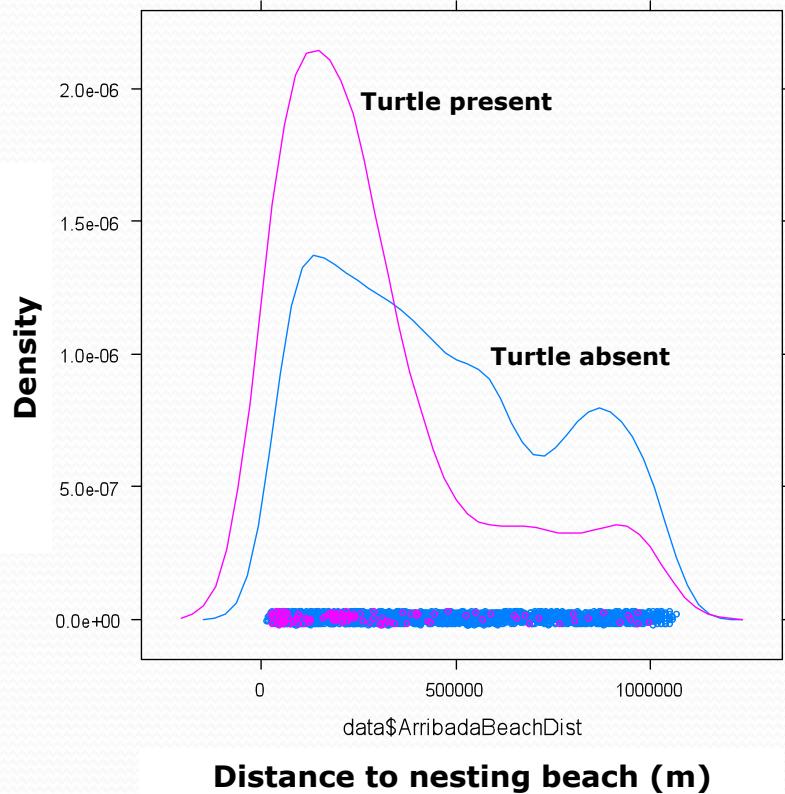
MGET statistics tools



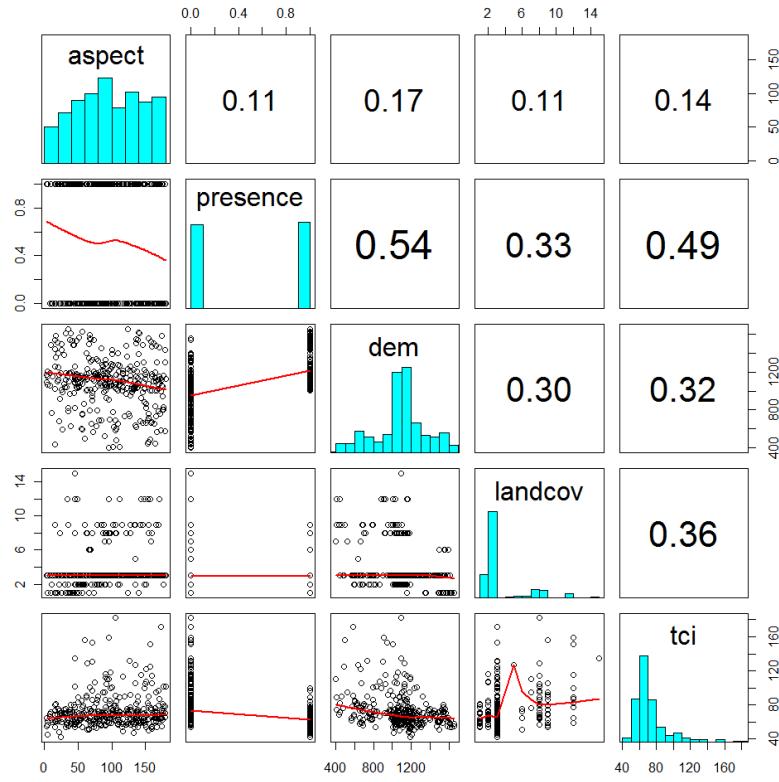
- Lots of tools, many more planned
- Built from Ben Best's ArcRStats / HabMod projects
- Tools require the R statistics program to be installed on your computer

Exploratory analysis

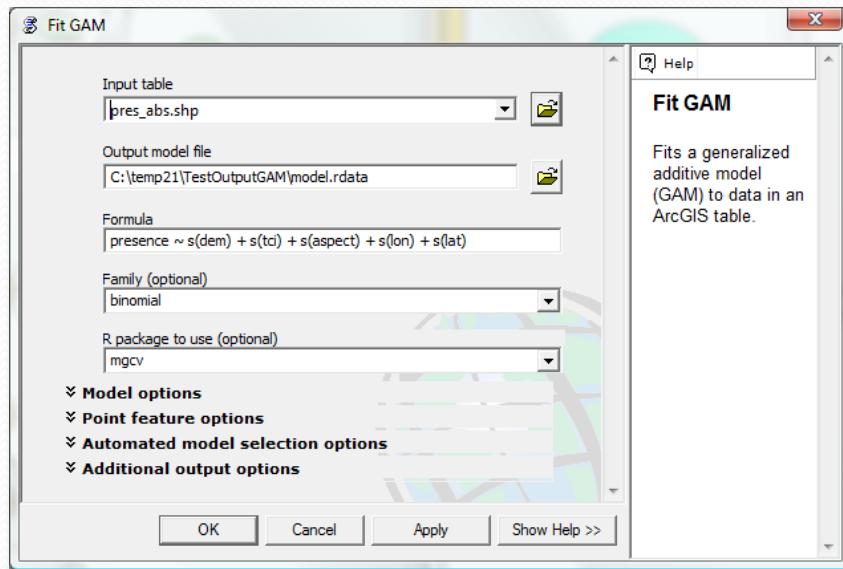
Density Histogram tool



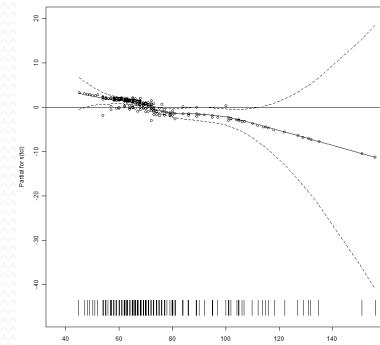
Scatterplot Matrix tool



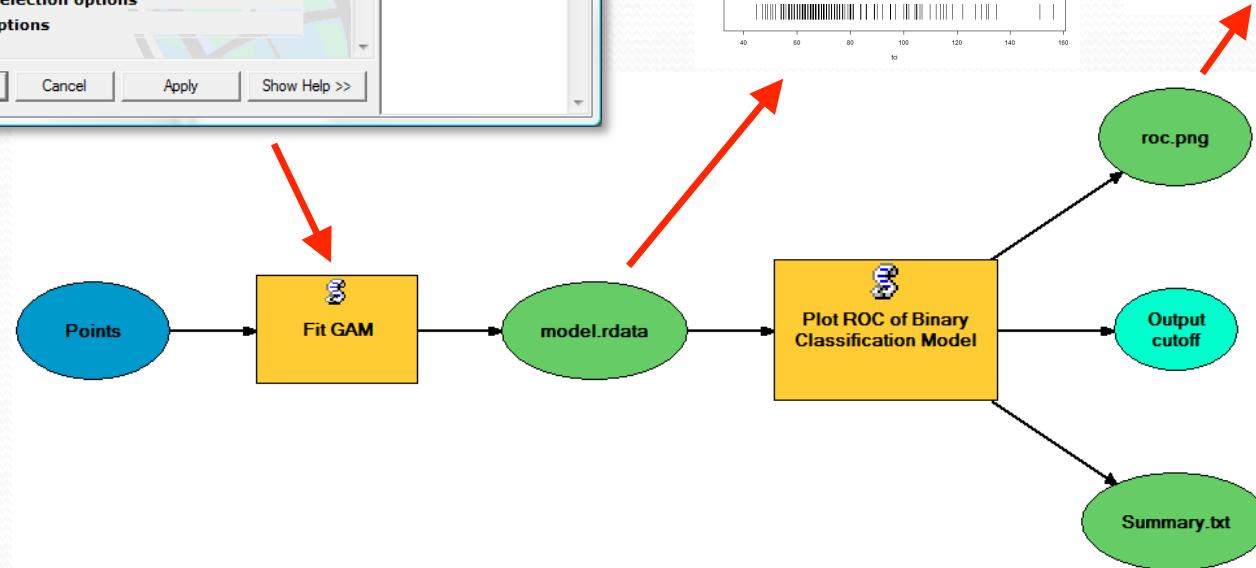
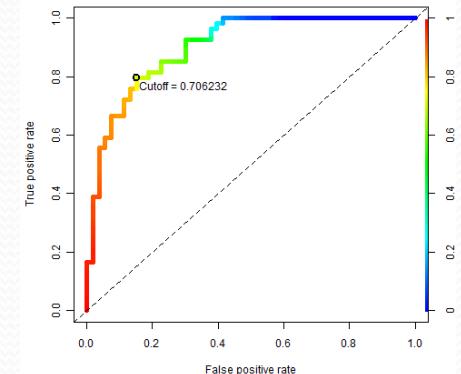
Fitting statistical models



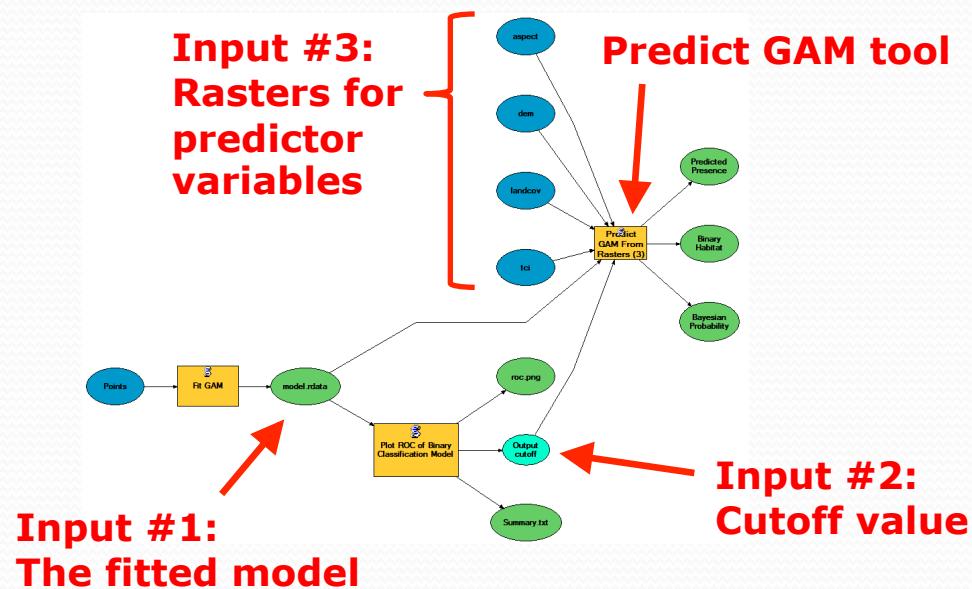
Term plots



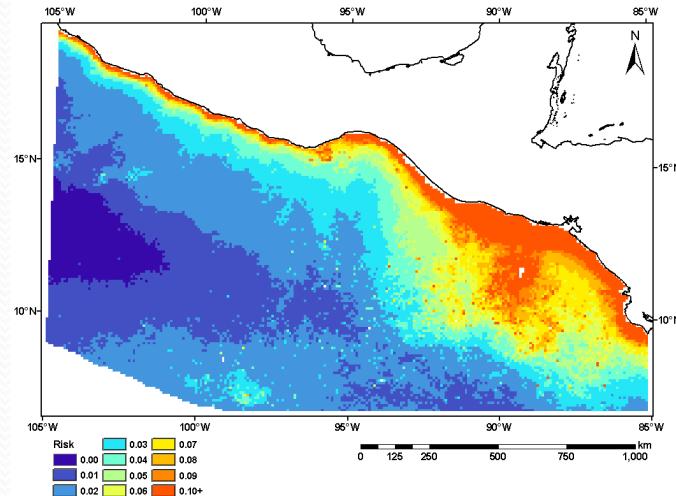
ROC plots



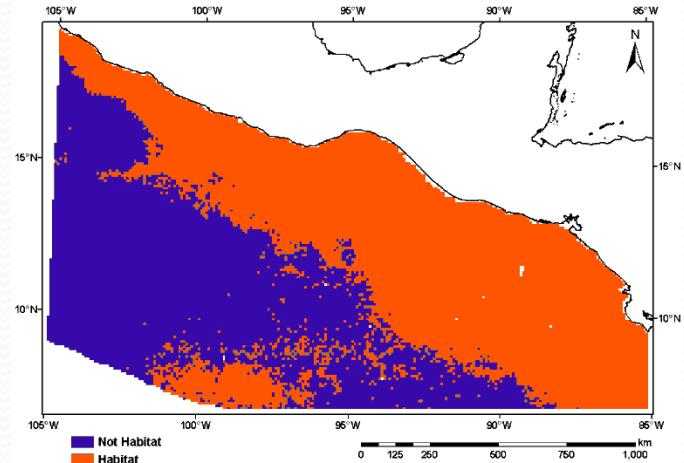
Predicting habitat maps from the model



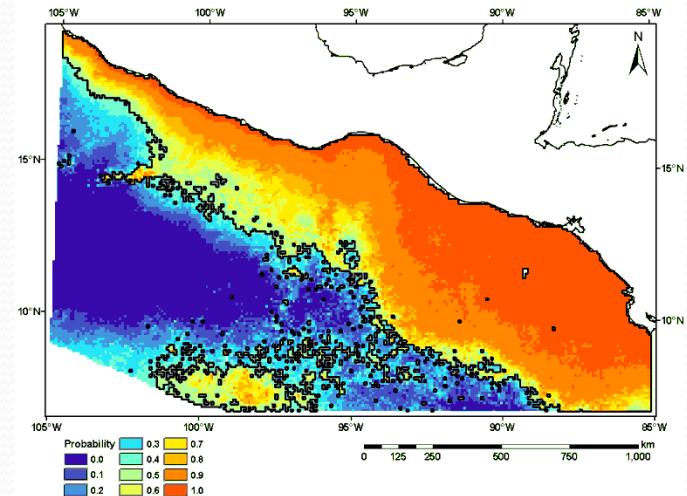
Predicted species presence

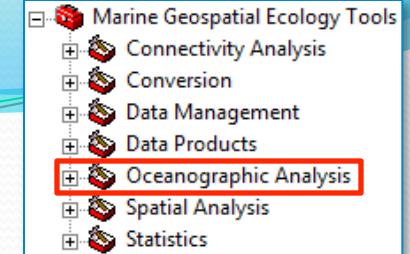


Binary habitat (cutoff = 0.025)



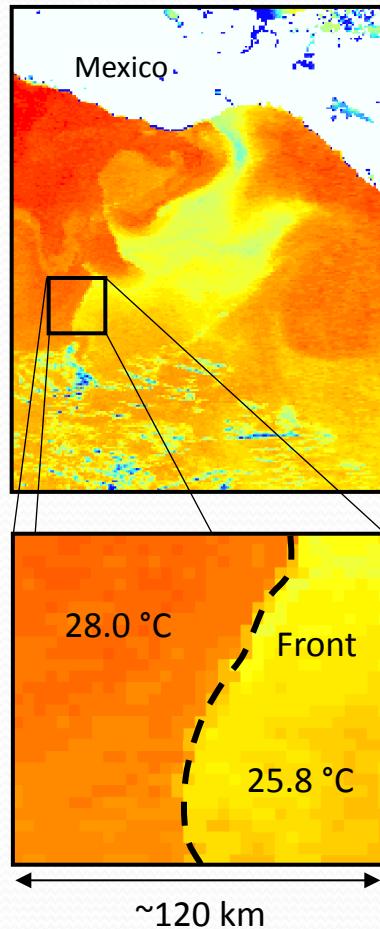
Bayesian probability that predicted presence ≥ 0.025





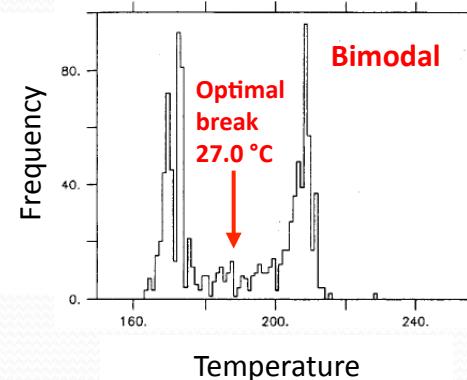
Identifying SST fronts

AVHRR Daytime SST
03-Jan-2005

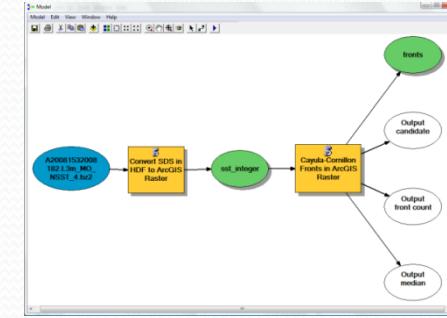


Cayula and Cornillion (1992) edge detection algorithm

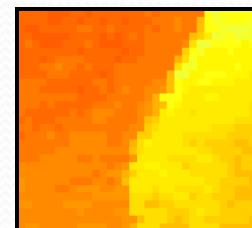
Step 1: Histogram analysis



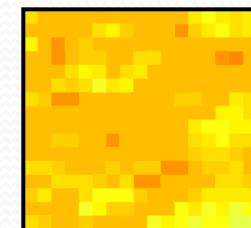
ArcGIS model



Step 2: Spatial cohesion test

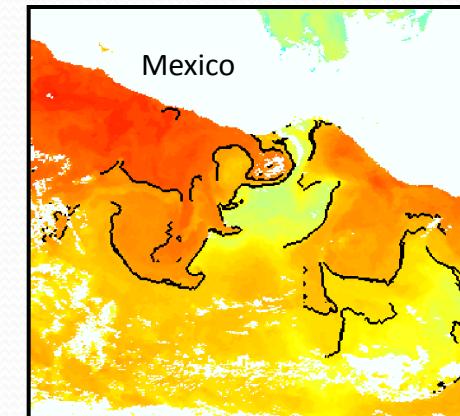


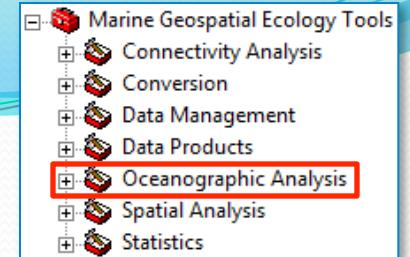
Strong cohesion →
front present



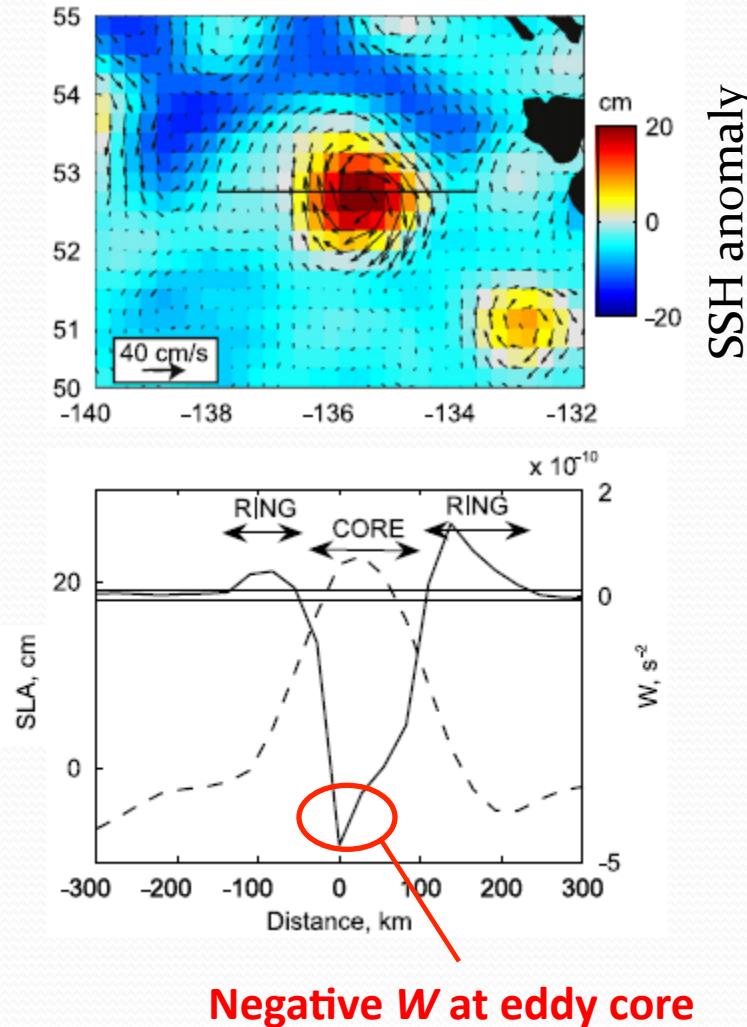
Weak cohesion →
no front

Example output





Identifying geostrophic eddies

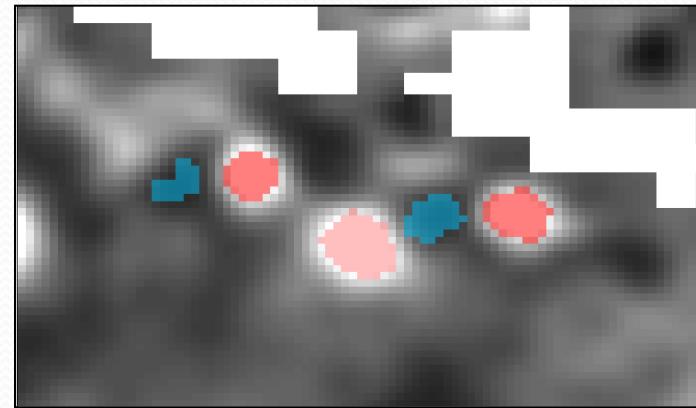


$$u = -\frac{g \partial h}{f \partial y}, \quad v = \frac{g \partial h}{f \partial x}.$$

$$\omega = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}, \quad s_n = \frac{\partial u}{\partial x} - \frac{\partial v}{\partial y}, \quad s_s = \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}.$$

$$W = s_n^2 + s_s^2 - \omega^2,$$

Example output



Aviso DT-MSLA 27-Jan-1993
Red: Anticyclonic Blue: Cyclonic

Available in
MGET 0.8

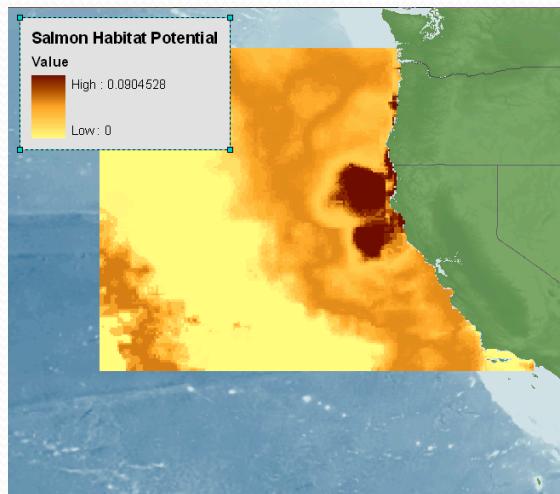
More examples (in class)

- Identifying salmon habitat based on SST
- Identifying Laysan Albatross generalized additive models
- Copy EDC-Salmon and unzip:
C:\arcgis\CCS\EDC-salmon



Example: Salmon & SST

- Identifying ideal habitat for Chinook Salmon using EDC and ArcGIS Spatial Analyst



Hinke, J.T., Foley, D.G., Wilson, C. & Watters, G.M. (2005) Persistent habitat use by Chinook salmon *Oncorhynchus tshawytscha* in the coastal ocean. *Marine Ecology Progress Series*, 304, 207–220.

Vol. 285: 181–192, 2005 MARINE ECOLOGY PROGRESS SERIES Mar Ecol Prog Ser Published January 19

Ocean habitat use in autumn by Chinook salmon in coastal waters of Oregon and California

Jefferson T. Hinke^{1,2,*}, George M. Watters², George W. Boehlert³, Paul Zedonis⁴

¹Joint Institute for Marine and Atmospheric Research, University of Hawaii, 1000 Pope Road, Honolulu, Hawaii 96882, USA

²NOAA Fisheries, Pacific Fisheries Environmental Laboratory, 1352 Lighthouse Ave, Pacific Grove, California 93950, USA

³Hatfield Marine Science Center, Oregon State University, 2300 S.E. Marine Science Dr, Newport, Oregon 97365, USA

⁴US Fish and Wildlife Service, 1655 Heindon Road, Arcata, California 95521, USA

ABSTRACT: Describing the ocean habitats used by Chinook salmon *Oncorhynchus tshawytscha* is an important step towards understanding how environmental conditions influence their population dynamics. We used data from archival tags that recorded time, temperature and pressure (depth) to define the coastal habitats used by Chinook near Oregon and California during the autumns of 2000, 2002 and 2003. We used a clustering algorithm to summarize the data set from each year and identified 4 general habitats that described the set of ocean conditions used by Chinook. The 4 habitats, defined primarily by depth and the time of day that these depths were occupied, were characterized as (1) shallow day, (2) shallow night, (3) deep and (4) deep. The definitions and use of each habitat were similar across years and the thermal characteristics of the habitats included water temperatures between 9 and 14°C. Thermal conditions in the habitats provided the best indicator of Chinook location in the coastal ocean. Chinook used 9 to 12°C water at least 52% of the time. Less than 10% of surface waters within the area where Chinook were released and recovered provided these temperatures. Cross sections of subsurface temperatures suggest that between 25 and 37% of the coastal water column was available to Chinook and contained water in the 9 to 12°C range. These results support hypotheses that link salmon-population dynamics to ocean temperatures. Continued monitoring of surface and subsurface thermal habitats may be useful for assessing the extent and quality of conditions most likely to sustain Chinook salmon populations.

KEY WORDS: Chinook salmon · Archival tag · California current · Essential fish habitat

Resale or republication not permitted without written consent of the publisher

INTRODUCTION

It is increasingly understood that both freshwater and marine habitats are critical to the maintenance of healthy Pacific salmon *Oncorhynchus* spp. populations. Understanding the linkages between environmental conditions, ocean habitats, survival and growth of these fishes is a primary goal of current salmon research initiatives (Boehlert 1997, Gargett 1997, Bisbal & McConaha 1998, Welch et al. 2003). Correlations of environmental conditions with catches (Mantua et al. 1997) and production (Bisbal et al. 1997, Cole 2000, Hobday & Boehlert 2001) have demonstrated a strong coupling of oceanographic variability

and salmon population dynamics. In particular, temperatures are correlated with salmon survival throughout much of the Pacific Ocean (Mantua et al. 1997, Cole 2000, Muster et al. 2002a). The effects of environmental conditions on population dynamics, however, are realized through the continuous interactions of individuals with their environment. Unfortunately, individual patterns of habitat use over time and the environmental conditions experienced in those habitats have rarely been measured.

Recognition of the importance of marine habitats for survival and production has fostered research aimed at identifying the environments that salmon actually experience (Boehlert 1997, Welch et al. 2003). Such

*Email: jefferson.hinke@noaa.gov

© Inter-Research 2005 · www.int-res.com

Acknowledgements

A special thanks to the many developers of the open source software that MGET is built upon! Also, folks that have helped with this talk:

Sara Maxwell, MGEL lab, Ecological Modeling workshop committee, and many others

Thanks to our funders:



For more information

Download MGET:

<http://code.env.duke.edu/projects/mget>

Email us:

jason.roberts@duke.edu, bbest@duke.edu,
elliott.hazen@duke.edu

Intro to habitat modeling:

Guisan, A., Zimmermann, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling* 135, 147–186.

Thanks for attending!

Autoregressive / Mixed models

- Autoregressive models incorporate a spatial covariance matrix (V_c) in the error term.
- Mixed models (GLMMs and GAMMs)
 - Can model random effects (e.g. tag deployment) and spatial autocorrelations in within-group errors for sequential data points.
 - Example (Hazen et al. 2009): Humpback whale surface feeding $\sim f(\text{environmental data, prey metrics}) + \text{random (whale)} + \text{AR1 correlation structure.}$

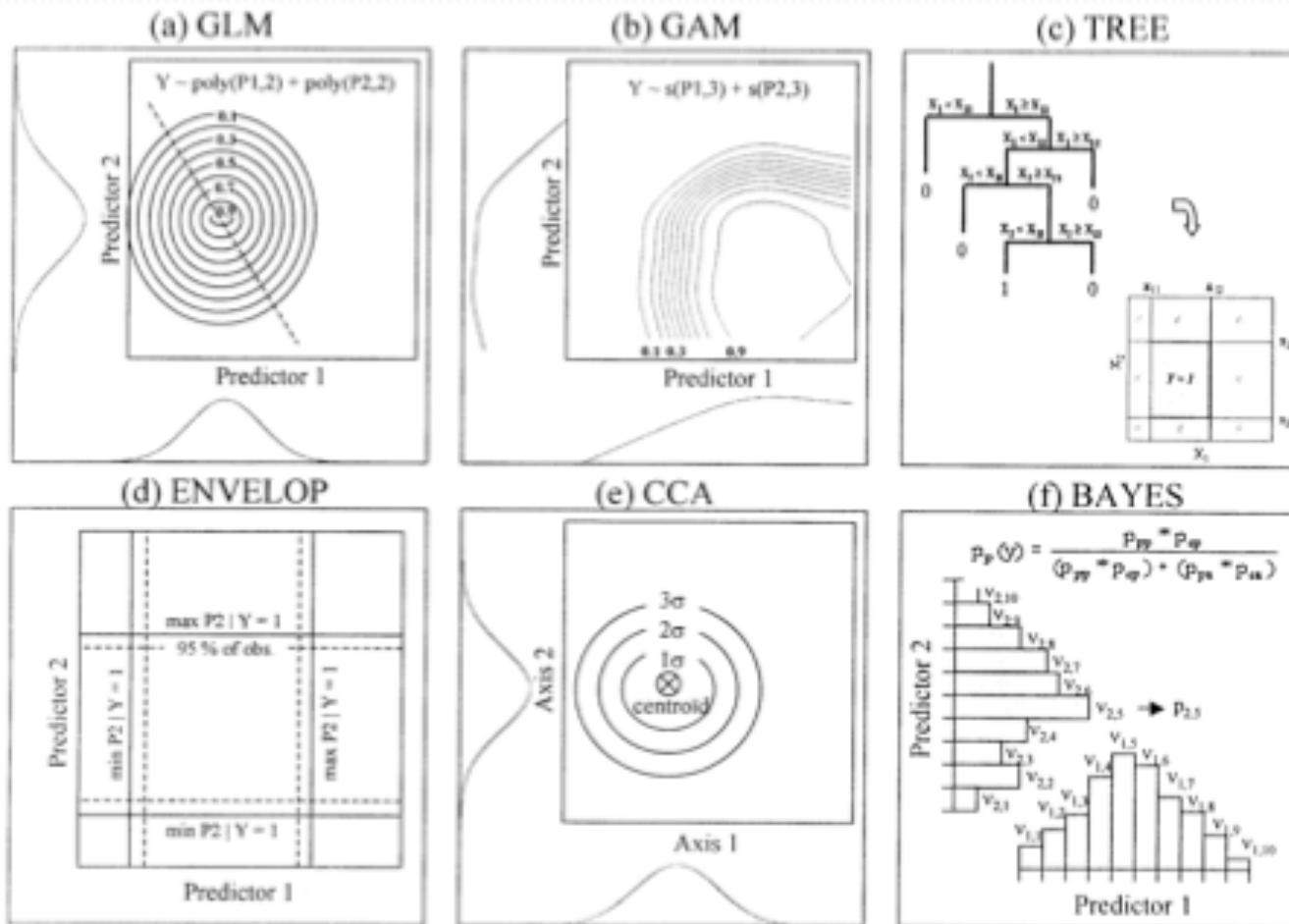
Data types

- Sightings – presence / absence, density
 - Binary response, zero inflated, many techniques
- Acoustic hydrophones – presence only
 - To be discussed later
- Tag data / focal follows – behavioral state/event
 - State models, movement models
- Vessels of opportunity – effort?
 - Presence only models

Predictor variables

- Location – bathymetry, distance from feature
- *In situ* oceanography / mooring / remotely sensed data
- Prey data – trawls, stomach contents, fisheries acoustics

Statistical Models



Guisan & Zimmermann (2000)

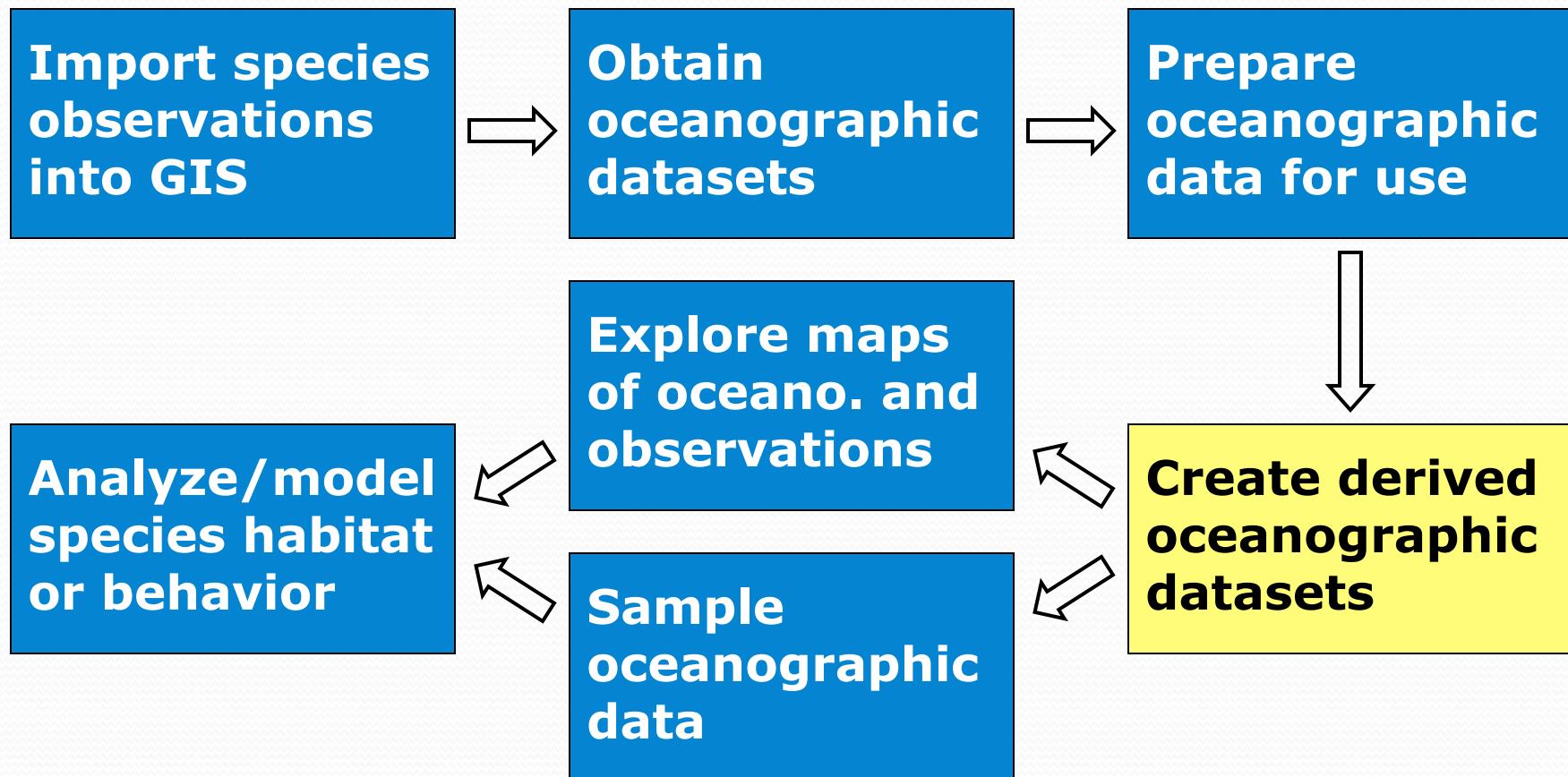
Correlative methods

- Linear vs. Additive models
 - Advantages to each
 - Assumptions – pseudo-absences,
 - Generalization - poisson distributed data (ZIP)

Introductions to the Software

- RTFM
- ArcGIS
 - Good, commercial help (+ video)
 - Training.ESRI.com
- Python
 - DiveIntoPython.org – free book
- R
 - A Beginner's Guide to R – free Springer book

Simplified workflow





Our Biases

- Disciplinary
 - Space (and Time)
 - Environment + Prey
 - Prediction
- Toolset
 - ArcGIS
 - Python
 - R