DATA MINING SPECIES DISTRIBUTION AND LANDCOVER

Dawn Magness – Kenai National Wildife Refuge

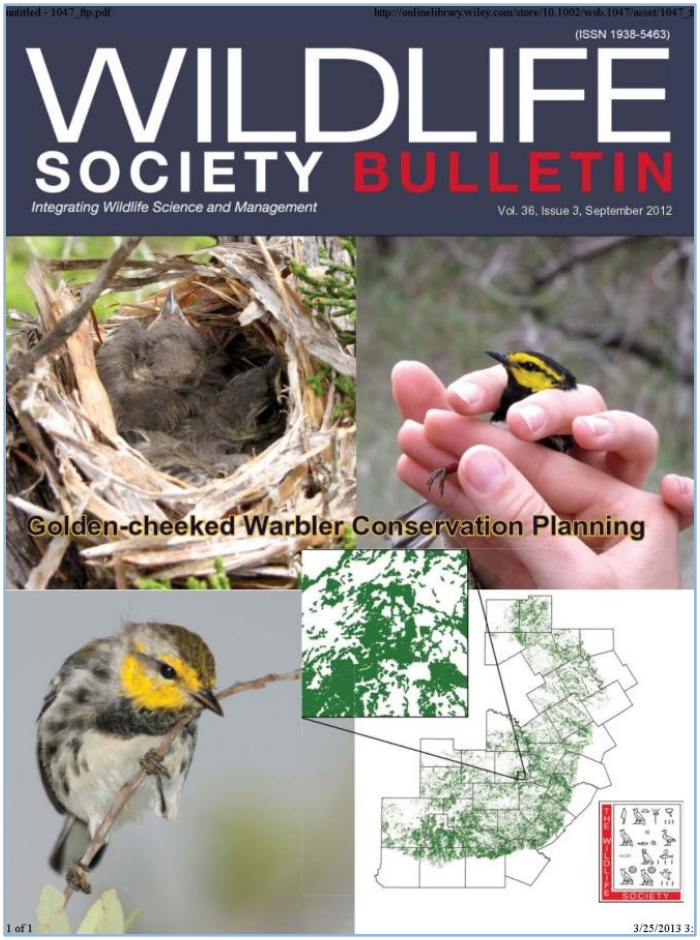
- **Why Data Mining**
- Random Forest Algorithm
- Examples from the Kenai
  - Species Distribution Model – Pattern
  - Landcover Model – Climate Envelope

I Just Want to Know Where Things Are

"resource managers generally lack detailed inventories and monitoring systems to provide them with an adequate baseline understanding of the plant and animal species that currently exist on the resources they manage"

- GAO 2007 Climate Change Report

Where Things Are Matters



The Federally Endangered Golden-cheeked Warbler: Nobel Intent, Inaccurate Assumptions

"Incomplete information on the distribution and abundance has led to substantial misunderstanding on species status and associated conservation goals"

- Morrison et al. 2012

Distribution Model Algorithms

Presence Only

- Gower Metric (DOMAIN)

Presence to Background

- Ecological Niche Factor Analysis (ENFA)

Presence to Absence (Or Pseudo-absence)

- GLM
- Boosted Regression Trees - Random Forest
- Genetic Algorithm (GARP)

Leo Breiman's Two Statistical Cultures

Data Modeling

- Assume stochastic data model in box; define data model *a priori*
- Estimate parameters
- Validation: Yes/No based on goodness of fit or residual analysis

Algorithmic Modeling (aka Data Mining)

- Function inside box complex and unknown
- Algorithm operates on x to predict response of y
- Validation: Predictive accuracy.

Brieman, L. 2001. Statistical modeling: the two cultures. Statistical Science. Vol 16: 199-31

Both Cultures Are Useful

Data Modeling

*Useful for **confirmatory analysis** and **hypothesis testing***

Algorithmic Modeling

*Useful for **accurate predictions** when there is **little prior knowledge***

House Finch Example

Compared Data model and Algorithmic Model

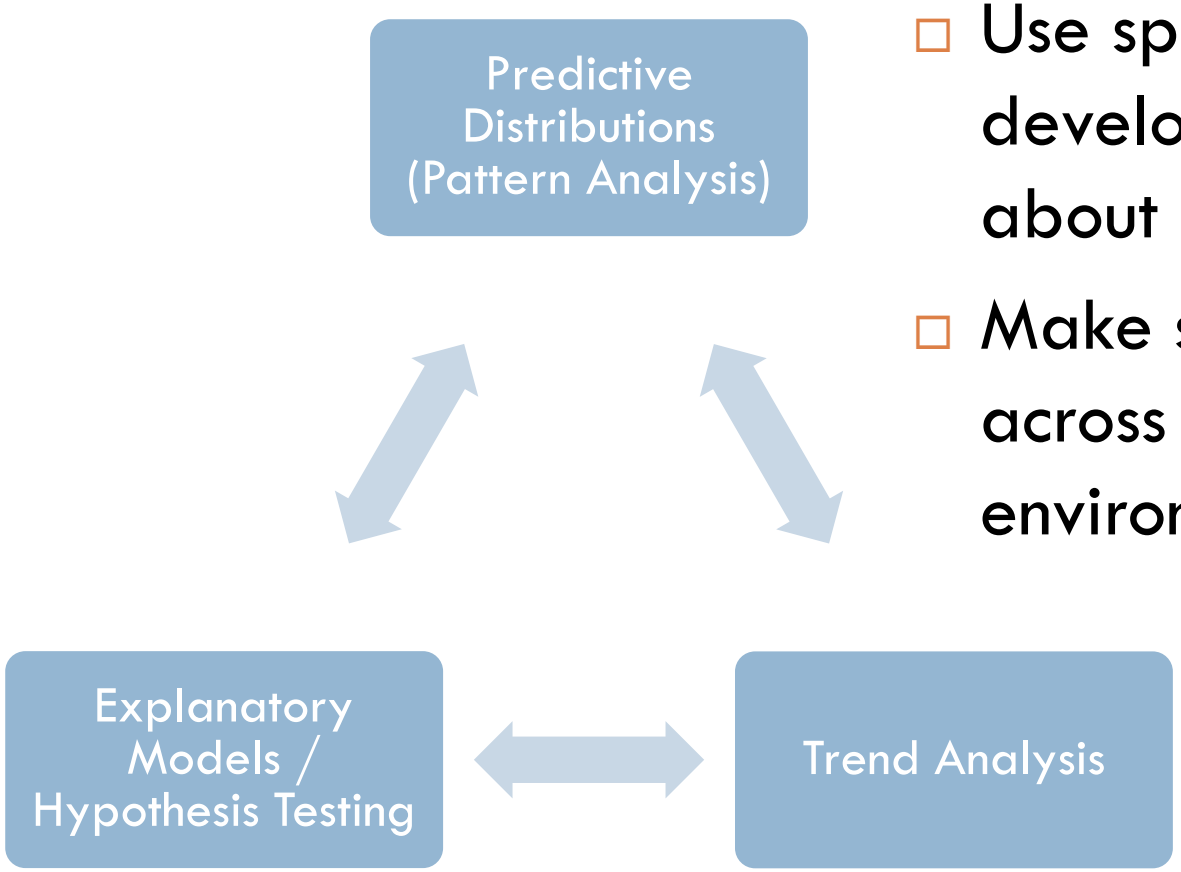
- Predictive Ability; slightly higher for algorithmic model
  - Data Mining: 205 predictors used
  - Data Model: 5 years of exploratory analysis
- Important Variables
  - 7 of 8 fixed effect predictors in data model identified as important by data mining

Hochachka, W. M. et al. 2006. Data-mining discovery of pattern and process in ecological systems. The Journal of Wildlife Management. 7: 2427 - 2437

Data Mining Really Good When

- Accurate prediction needed
- When little prior knowledge about system exists
  - Handles many predictors, correlated predictors alright
  - Complex Systems: Weak signals, multiple interactions, and non-linear relationships are handled well

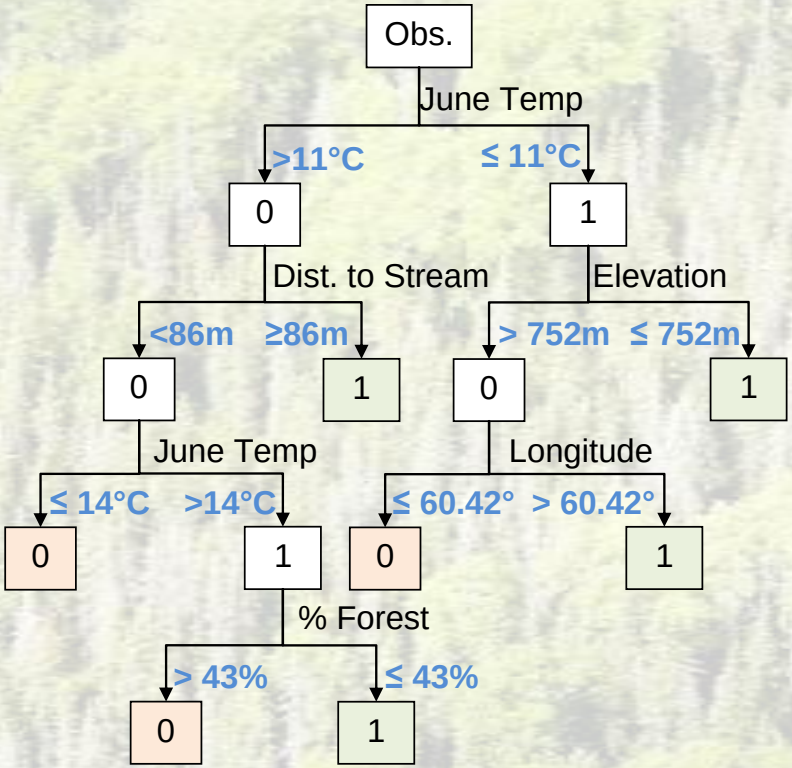
Leveraging Where Things Are



- Use spatial patterns to develop hypothesis about mechanisms
- Make sure to sample across the range of environmental conditions

- Why Data Mining
- **Random Forest Algorithm**
- Examples from the Kenai
  - Species Distribution Model – Pattern
  - Landcover Model – Climate Envelope

Random Forest Algorithm – Many Trees



- **Boosting** - *successive trees give extra weight to points incorrectly predicted by earlier trees*
- **Bagging** – *successive trees do not depend on earlier trees; each is independently constructed using a bootstrap sample of data set*

Random Forest Algorithm Steps

- Draw n_{tree} bootstrap sample from original data
- For bootstrap sample, grow an unpruned tree. At each node, randomly select n = m_{try} predictors choose best split from this subset
- Predict new data by aggregating predictions of n_{tree} trees. (majority votes for classification; average for regression)
- Class weights

Liaw, A. And W. Wiener et al. 2002. Classification and regression by randomForest. R News. ISSN 1609-363. Vol. 2/3: 18- 22.

Random Forest Error

- At each bootstrap iteration, predict data not in the bootstrap sample (the "out-of-bag" or OOB data)
- Aggregate OOB predictions. Calculate the error rate and call it the OOB estimate of error

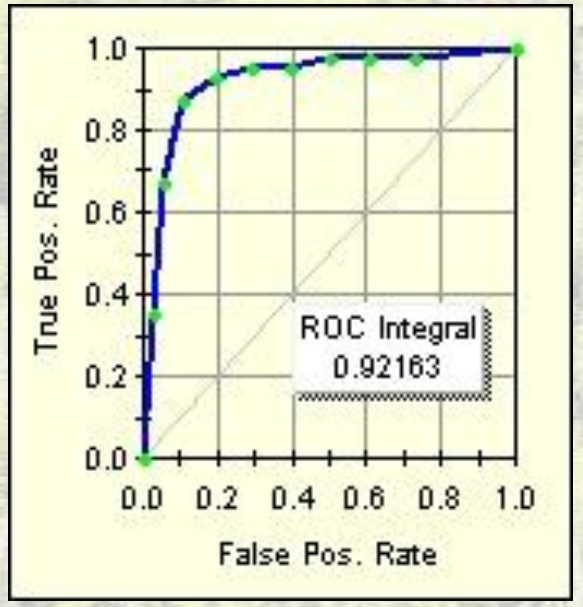
Liaw, A. And W. Wiener et al. 2002. Classification and regression by randomForest. R News. ISSN 1609-363. Vol. 2/3: 18- 22.

Random Forest Variable Importance

- Estimates and ranks importance of a variable by looking at how much prediction error increases when data for that variable is permuted while all others are left unchanged.

Liaw, A. And W. Wiener et al. 2002. Classification and regression by randomForest. R News. ISSN 1609-363. Vol. 2/3: 18- 22.

Measures of Predictive Ability

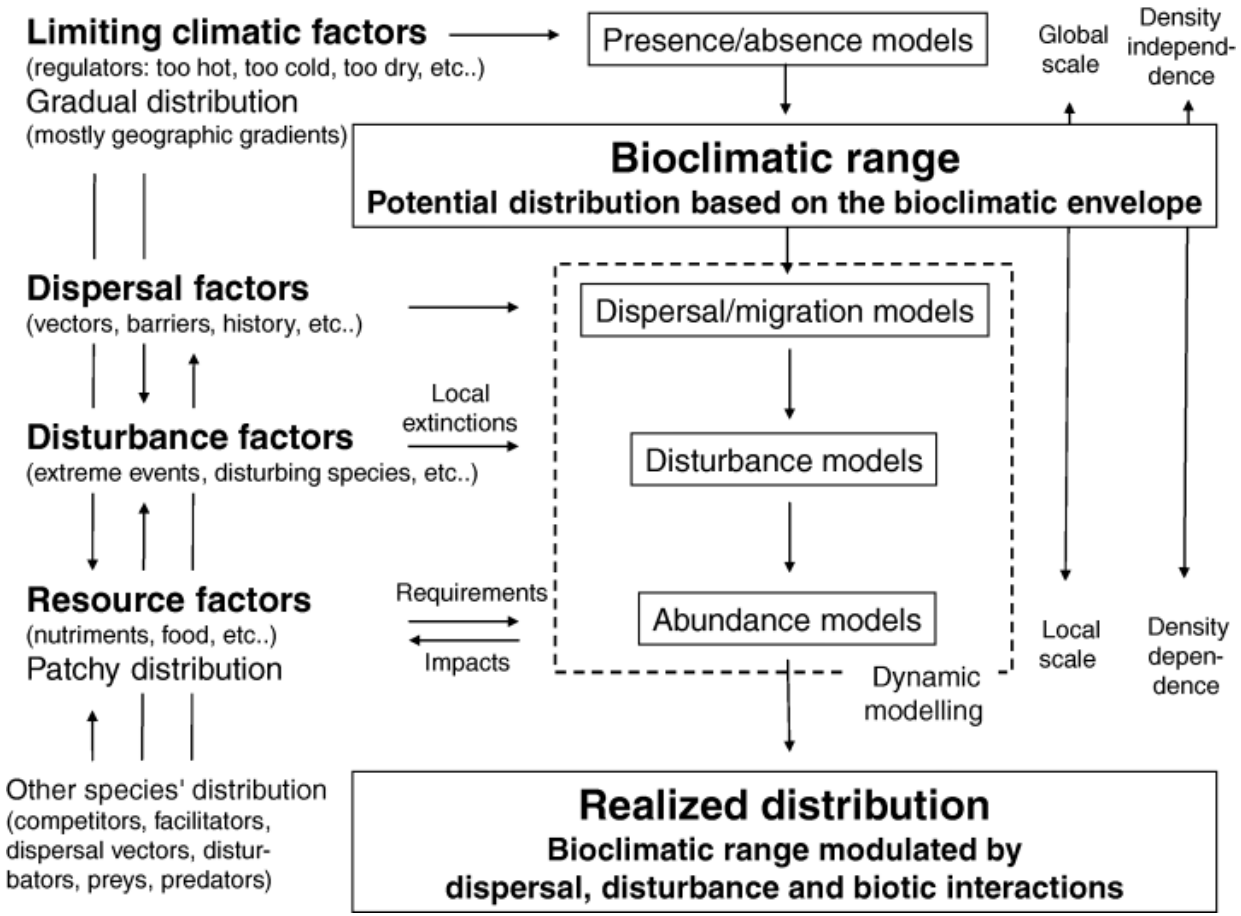


Predictive Ability

Actual Class	Out of Bag Prediction 0	Out of Bag Prediction 1	Percent Correct
0	190	25	**88%**
1	5	35	**88%**

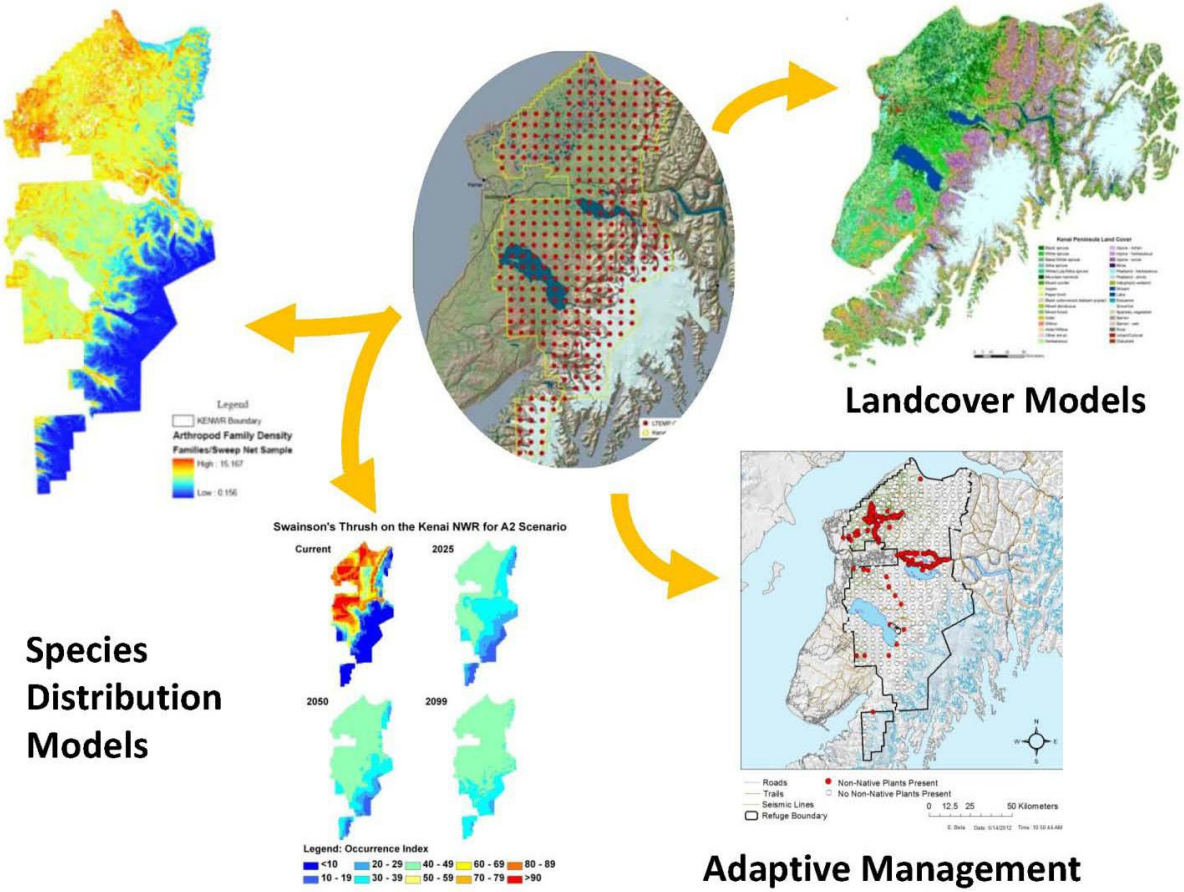
- Why Data Mining
- Random Forest Algorithm
- **Examples from the Kenai**
  - **Species Distribution Model – Mining Pattern**
  - Landcover Model – Climate Envelope

Built on Niche Theory



Guisan, A. and W. Thuiller. 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters. 8” 993 - 1009

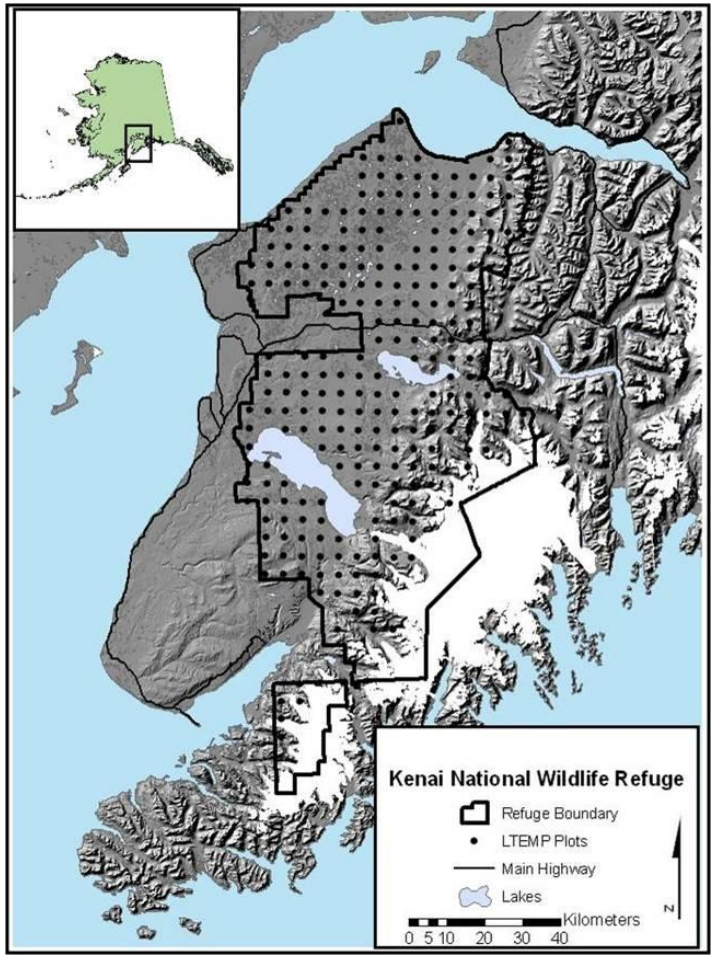
Long Term Ecological Monitoring Program on Kenai National Wildlife Refuge



Morton, J. M. et al. *In Press*. An integrative approach to inventory, monitoring, and modeling species diversity in a changing climate. Journal of Fish and Wildlife Management

Kenai NWR Monitoring Program

- 255 Plots with Variable Radius Point Count
- Avoids Some Pitfalls
  - Spatial Bias
  - False Positives from Gross Misidentification
  - Negatives (although with detection error)

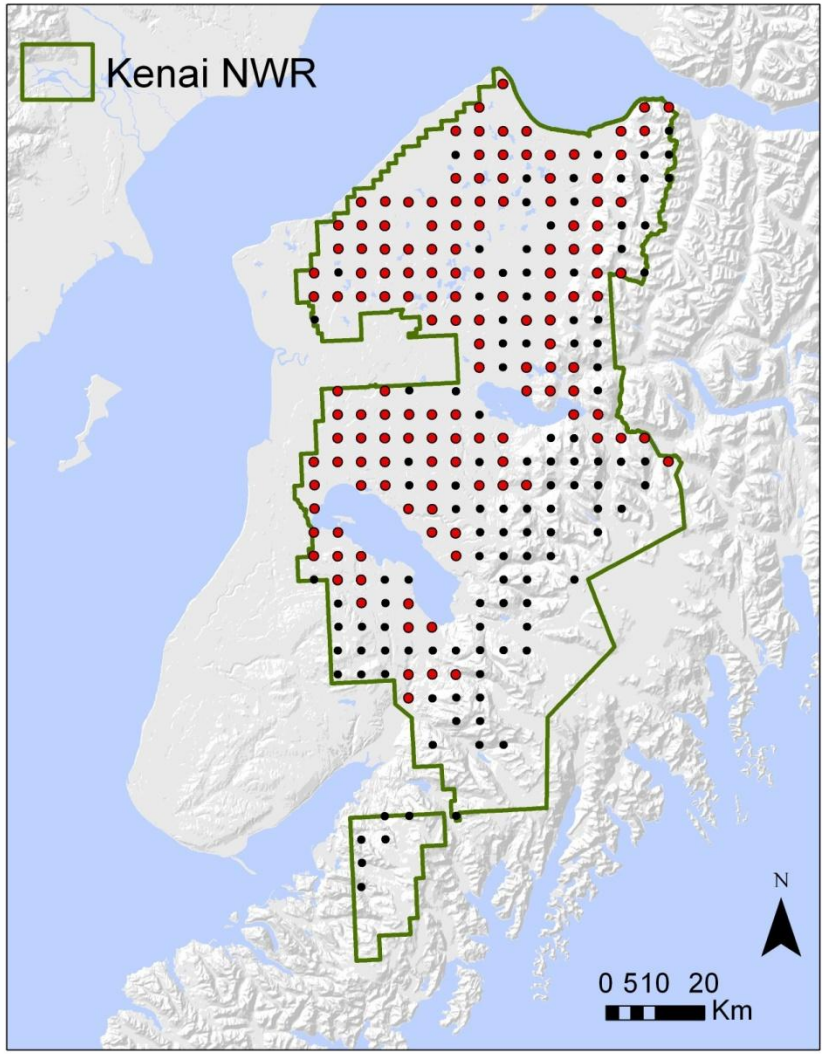


Kenai Bird Distribution for Monitoring

- Built Random Forest models for **40 species** present within 200m of sampling points
- Common set of **157 predictor variables** available as GIS layers
  - Topographic features
  - Spatial Structure
  - Climate
  - Anthropogenic
  - Vegetation
  - 5 Nonsense

Magness, D.R. et al. 2008. Using Random Forests to provide predicted species distribution maps as a metric for ecological inventory & monitoring programs. Pages 209-229 *in* Applications of Computational Intelligence in Biology: Current Trends and Open Problems. Studies in Computational Intelligence, Vol. 122, Springer-Verlag

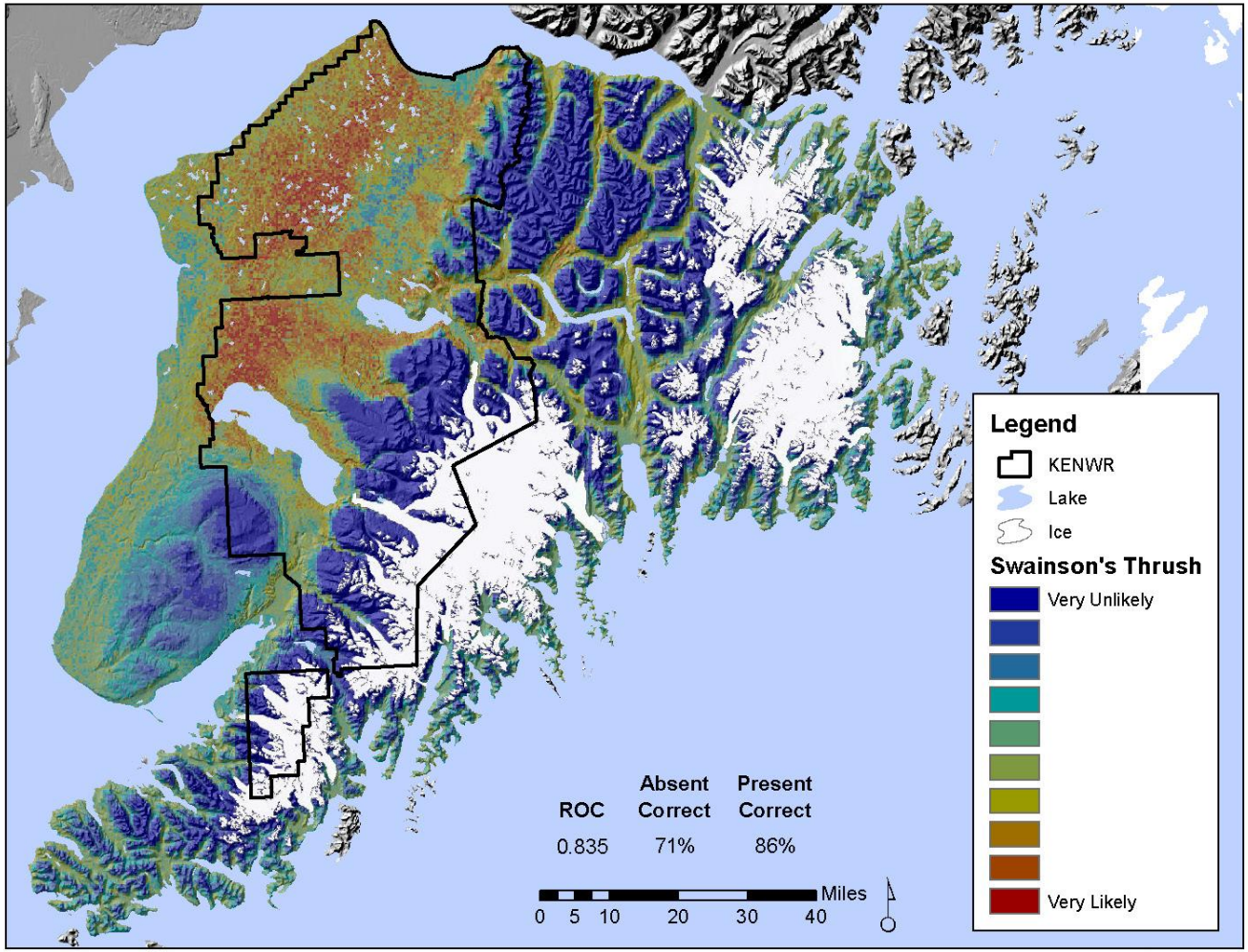
Detected 80 species on grid (of 96 known to breed on the Kenai Peninsula)

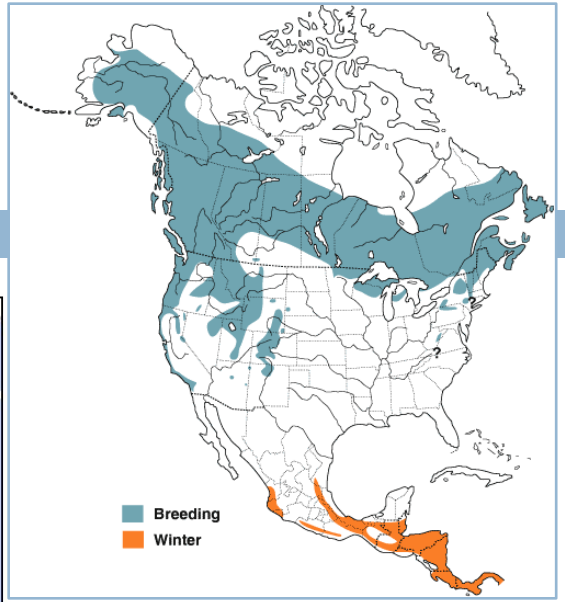


40 birds present at >3% of sites for modelling

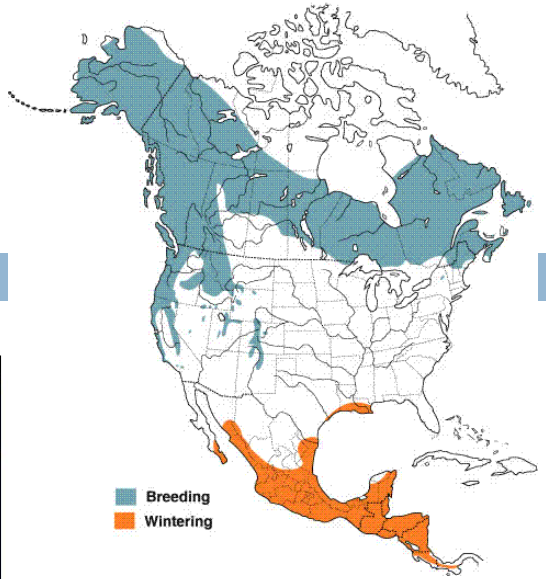
Swainson's Thrush
present at 131 sites (51%)

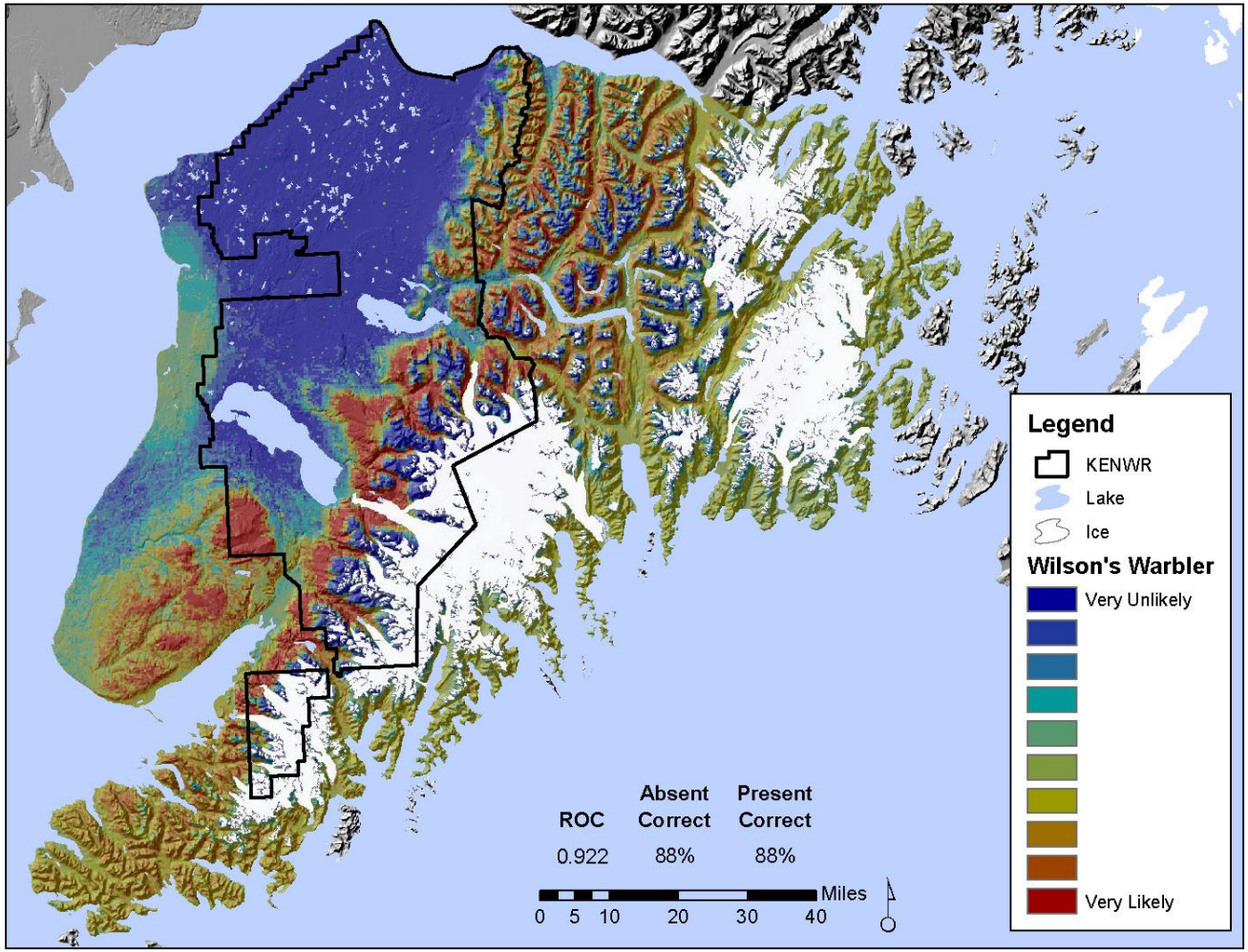
Surface Response





Wilson's Warbler





Predictive Ability

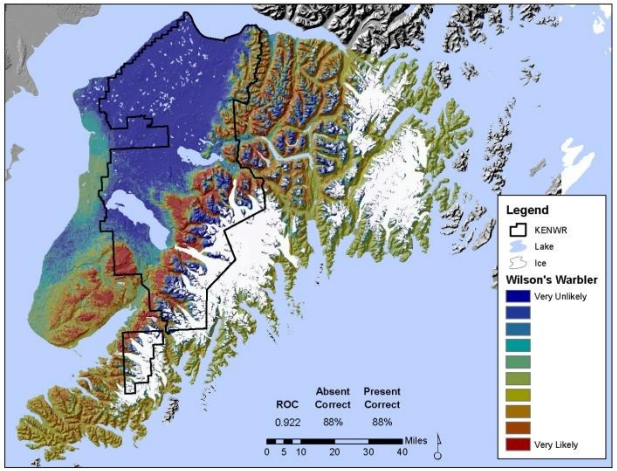
Actual Class	Out of Bag Prediction 0	Out of Bag Prediction 1	Percent Correct
0	190	25	88%
1	5	35	88%

Variable Importance	Score
BIOC18	100
DEM	99.6
BIOC12	96.5
TMAX05	89.9
TMAX06	89.1
BIOC17	88.9
TMIN05	80.72

Wilson's Warbler

Predictive Ability

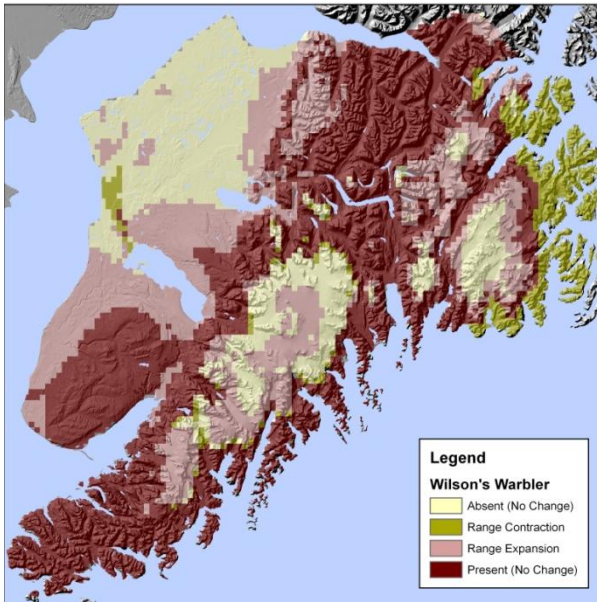
Actual Class	Out of Bag Prediction 0	Out of Bag Prediction 1	Percent Correct
0	190	25	88%
1	5	35	88%



t=1

t=2 in 2014.

Change detection using predictive maps with error?

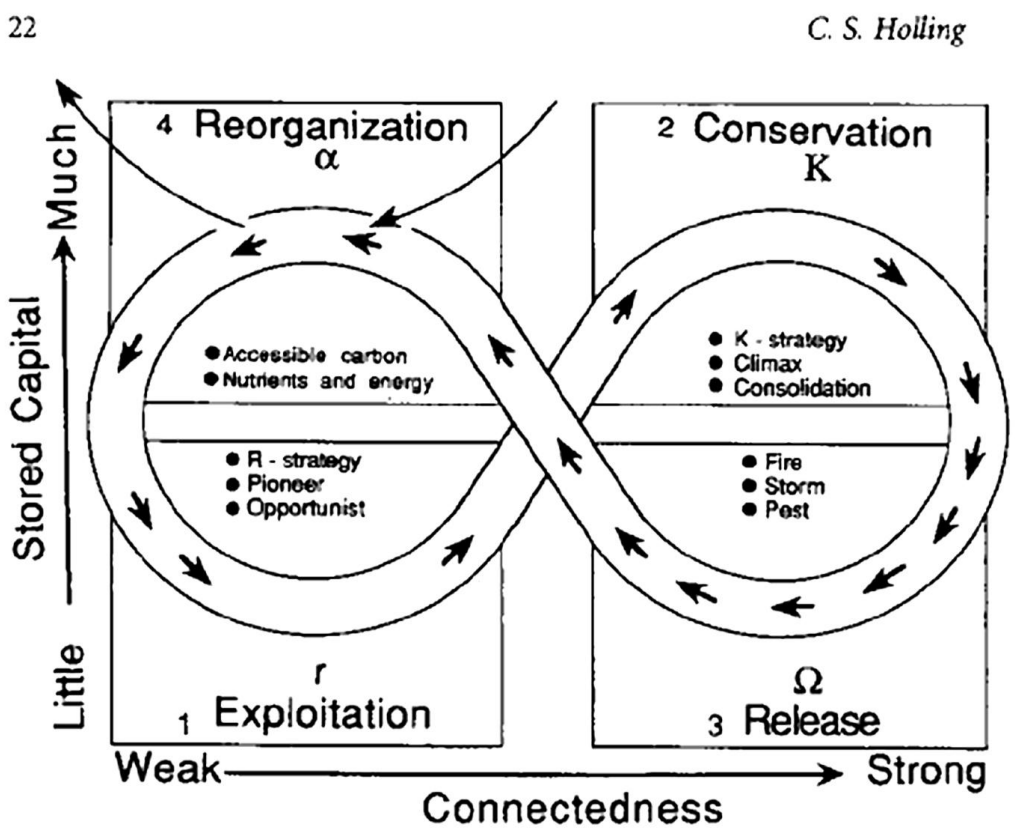


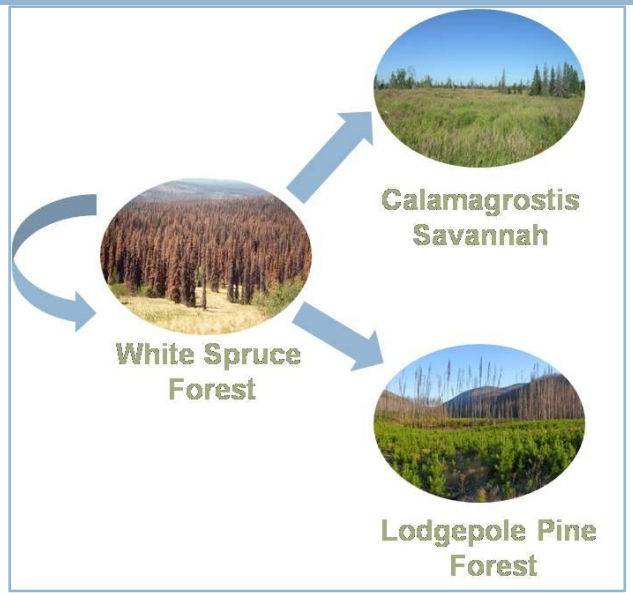
2099 climate-envelope model

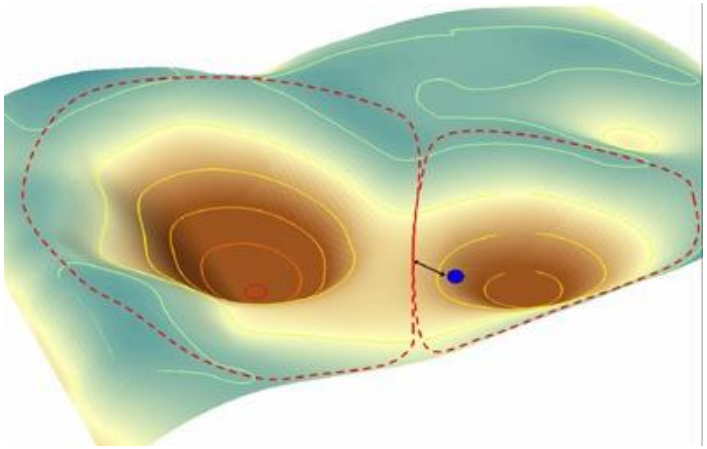
- Why Data Mining
- Random Forest Algorithm
- **Examples from the Kenai**
  - Species Distribution Model – Pattern
  - **Landcover Model – Climate Envelope**

Magness and Morton. Accepted. Journal of Fish and Wildlife Management

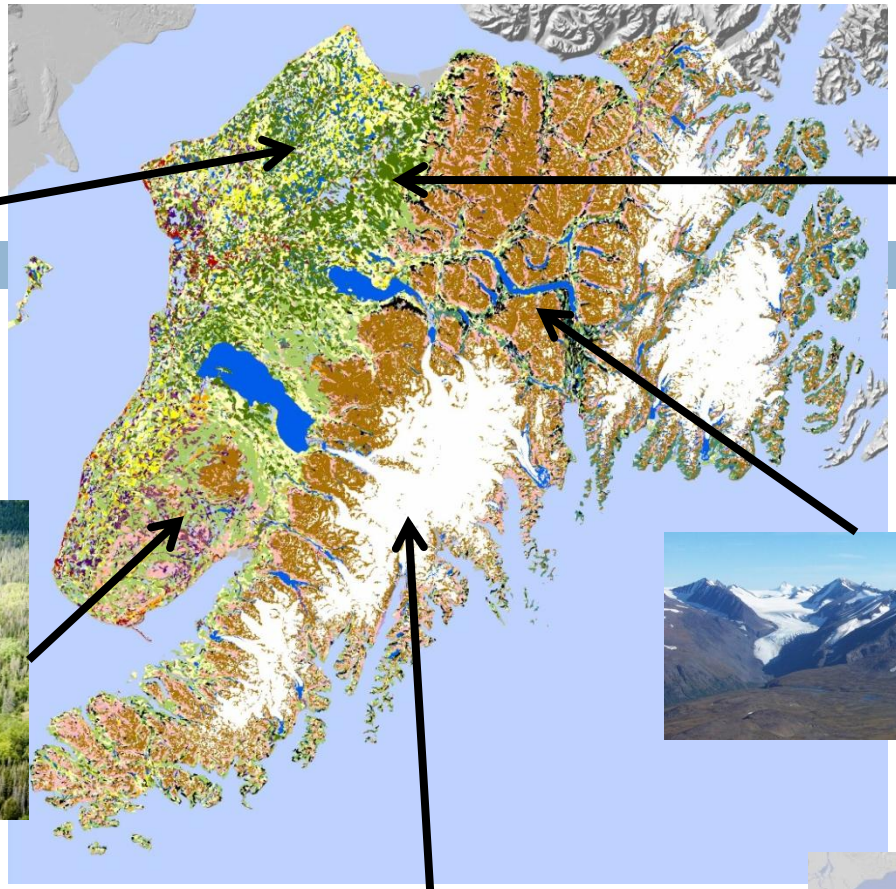
Resilience Theory



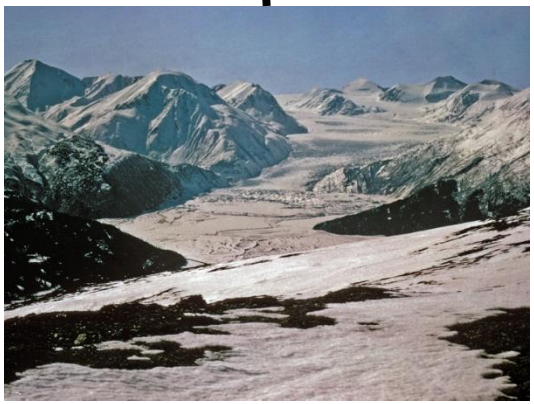


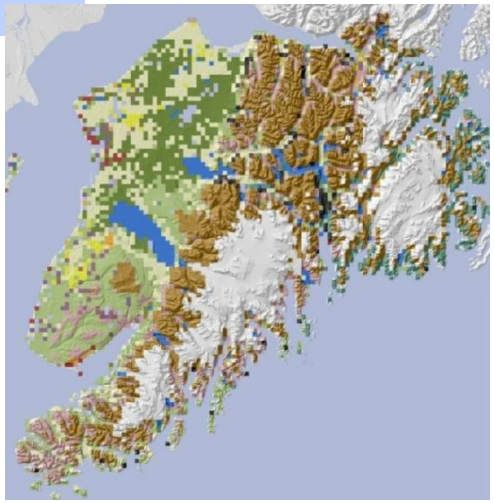


The Resilience Alliance (resiliencealliance.org)



- Alder
- Alpine
- Anthropogenic
- Birch
- Black Spruce
- Herbaceous
- Mixed Conifer
- Mixed Forest
- Mountain Hemlock
- Other Deciduous
- Other Shrub
- Snow or Ice
- Water
- Wetland Graminoid
- Wetland Shrub
- White or Lutz or Sitka Spruce
- Willow





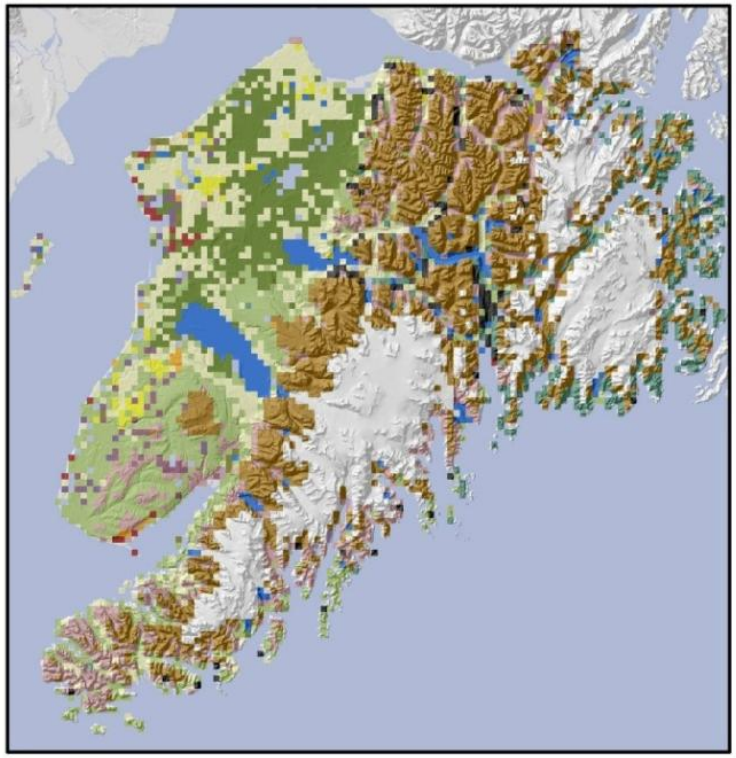
Climate-envelope for Landcover

- Used Random Forest to predict landcover using 28 climate variables
- Forecast to 2039, 2069, and 2099 timestep using Transition Decisions Rules
  - Maximum Wins
  - Transition when below 0.75
  - Transition when below 0.5

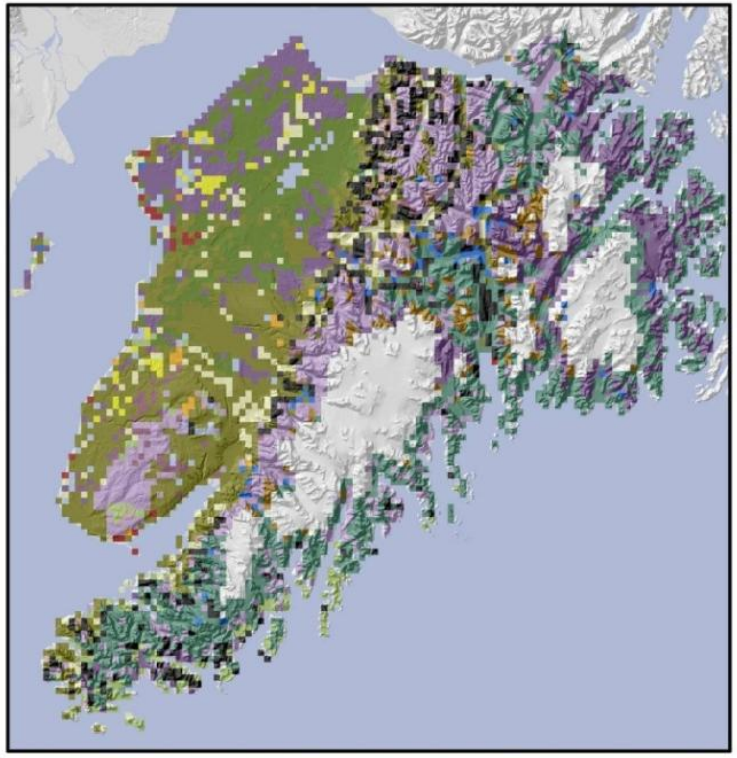
Scenarios Forecast (n=24)

- SRES Scenarios
  - a2
  - a1b
  - B1
- GCMs (n=6)
- Transition Decisions Rules
  - Maximum Wins
  - Transition when below 0.75
  - Transition when below 0.5

2009



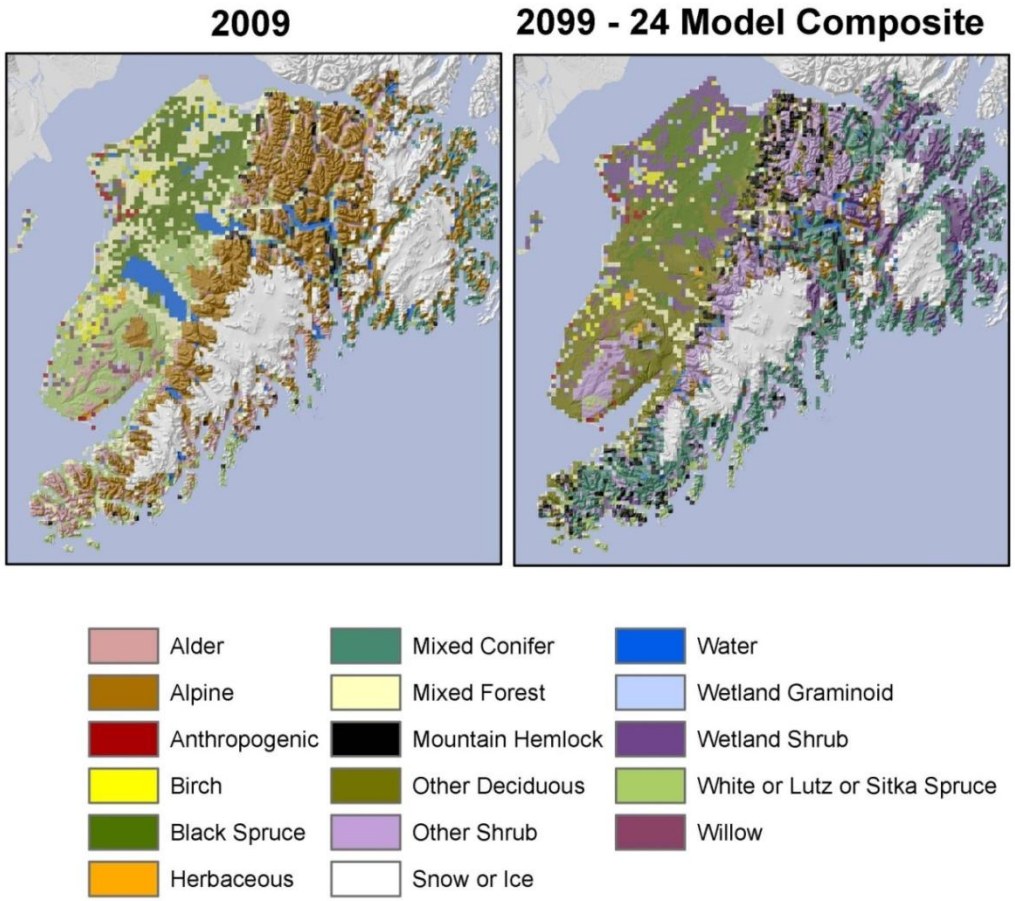
2099 - 24 Model Composite

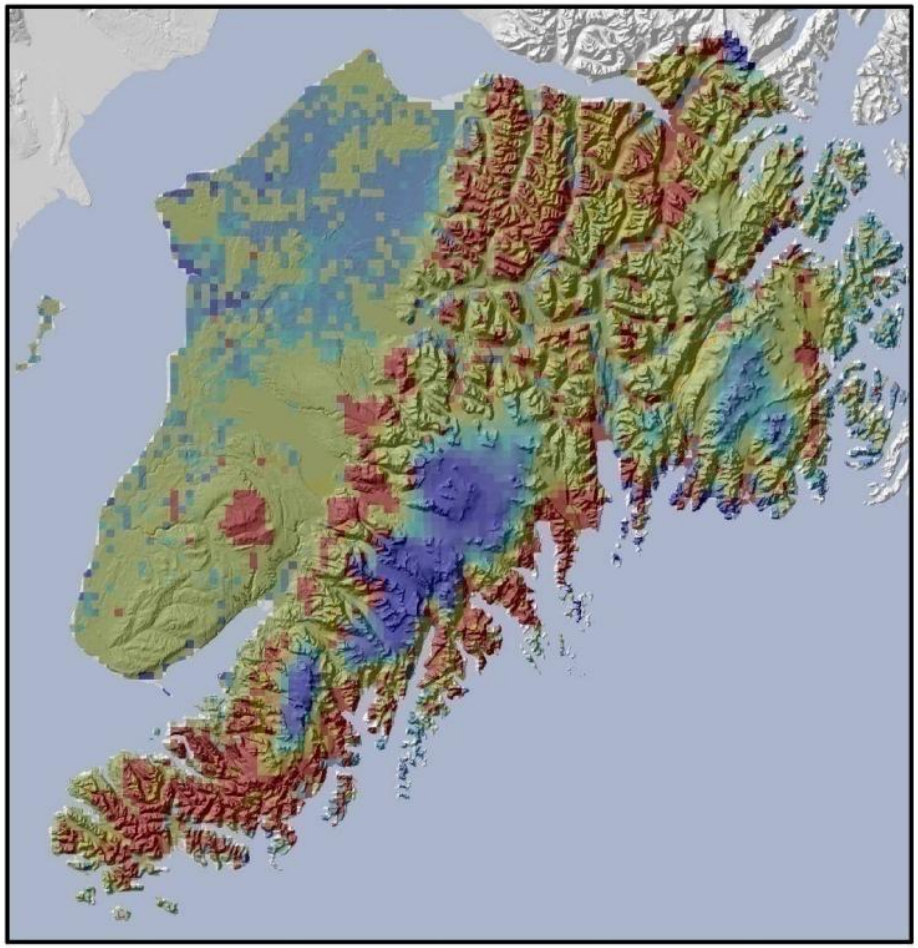


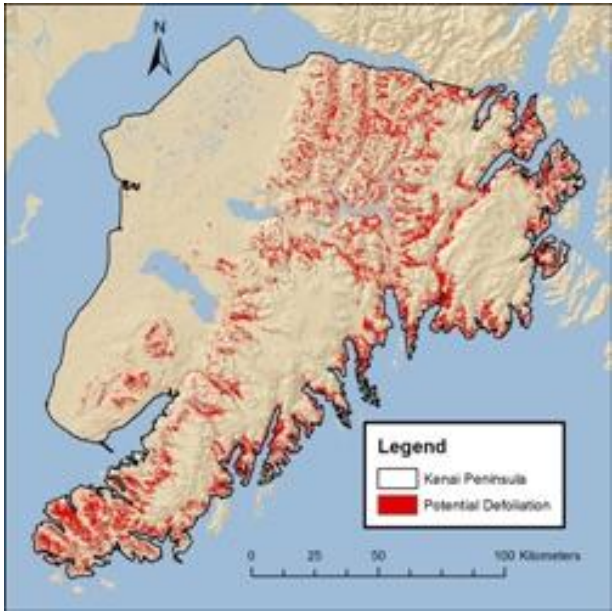
- Alder
- Alpine
- Anthropogenic
- Birch
- Black Spruce
- Herbaceous
- Mixed Conifer
- Mixed Forest
- Mountain Hemlock
- Other Deciduous
- Other Shrub
- Snow or Ice
- Water
- Wetland Graminoid
- Wetland Shrub
- White or Lutz or Sitka Spruce
- Willow

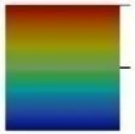
Beyond Forecast Map

- Identify potential spatial refugia and transitional areas?
- Identify robust trends?









Transition - All 24 Models

Refugia - All 24 Models

Trends in Acres

Landcover Type	Acres 2009	Min. Acres	Max. Acres	Trend
Alder	**609,843**	**0**	**439,838**	**-**
Alpine	**1,462,832**	**0**	**961,713**	**-**
Birch	61,281	0	162,098	
Black Spruce	477,397	0	484,316	
Herbaceous	16,803	0	49,420	
Mixed Conifer	**281,694**	**507,049**	**1,912,554**	**+**
Mixed Forest	**676,066**	**0**	**635,541**	**-**
Mountain Hemlock	100,817	13,838	650,367	
Other Deciduous	**29,652**	**43,490**	**1,997,556**	**+**
Other Shrub	**19,768**	**261,926**	**1,292,827**	**+**
Snow or Ice	**1,375,853**	**105,759**	**1,269,106**	**-**
Water	223,378	0	282,682	
Wetland Graminoid	104,770	21,745	598,970	
Wetland Shrub	**24,710**	**24,710**	**2,357,334**	**+**
White or Lutz or Sitka Spruce	739,323	0	741,300	
Willow	44,478	0	208,552	
Novel	0	0	347,917	

Robust Trends With Evidence

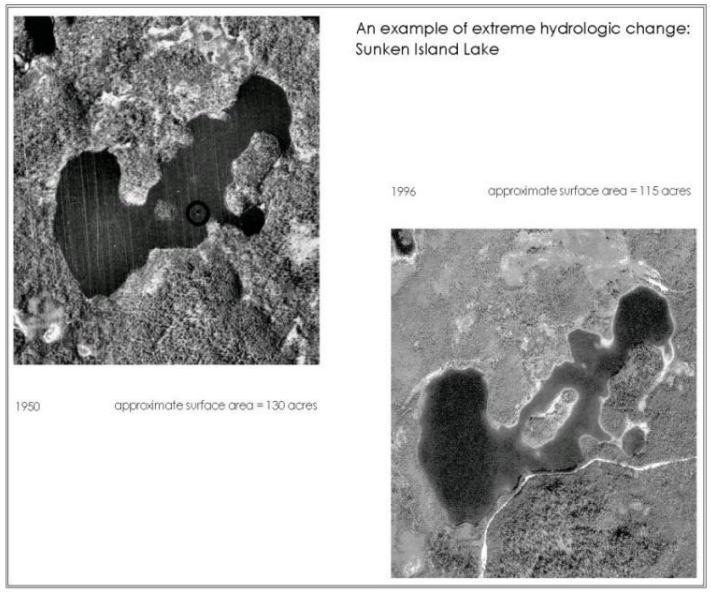
- Alpine Tundra Decreases or Lost; Mixed Conifer Conversion in Eastern Kenai Mountains
  - Treeline rise of 10m/decade from1950s
  - Observed defoliation and die off
- Perennial Snow and Ice Decreases
  - Harding Ice Field lost 5% of surface area since from 1950-1985

Robust Trends With Evidence

- Increase of Other Deciduous; Usually Replacing White Spruce
  - No evidence for increased Cottonwood
  - White Spruce Beetle Kill Event
  - Increasing grassland cover; Low spruce recruitment
  - Spring, grass carried fire

Robust Trends With Evidence

- Black Spruce Remains in Kenai Lowlands
  - Black spruce colonizing peatland during period of warming temperatures and increased evapotranspiration
  - Fire modeling also suggests coniferous forests remain
- Mixed Forest Decline
  - White Spruce Thinning, but otherwise no evidence
- Alder Stand Decline
  - Defoliation, but otherwise no evidence

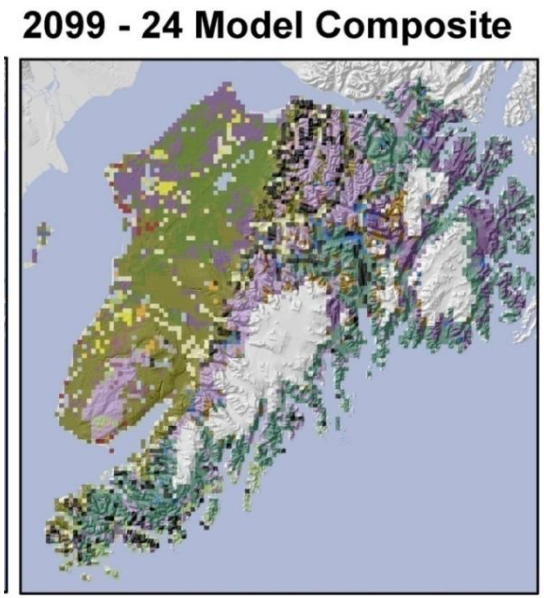


Signal Convergence Reduces Uncertainty

Anecdotal Observations

Change Detection Using Remote Sensing

Grid-based Monitoring



Adaptive Management

Targeted Research to Understand Mechanisms Driving Patterns

Mechanistic Models

Predictive Models

Thanks To:

Salford Systems

EWHALE Lab – University of Alaska

SNAP