SOMOSPIE

SOil MOisture SPatial Inference Engine

Danny Rorabaugh, Mario Guevara, Ricardo Llamas,
Joy Kitson, Rodrigo Vargas, Michela Taufer

UT Booth @ SC19

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

BIG ORANGE. BIG IDEAS.®

Acknowledgements

M. Guevara

T. Johnston

J. Kitson

R. Llamas

P. Olaya

E. Racca

A. Schwartz

K. Suarez

M. Taufer

R. Vargas

Sponsors:

NSF

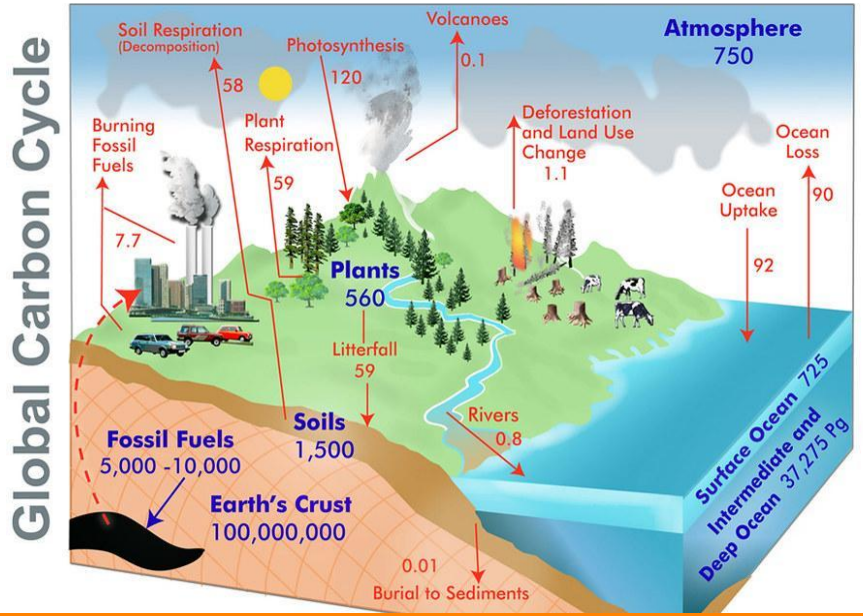
XSEDE

Jetstream

Soil moisture?

Why soil moisture?

Environmental Sciences

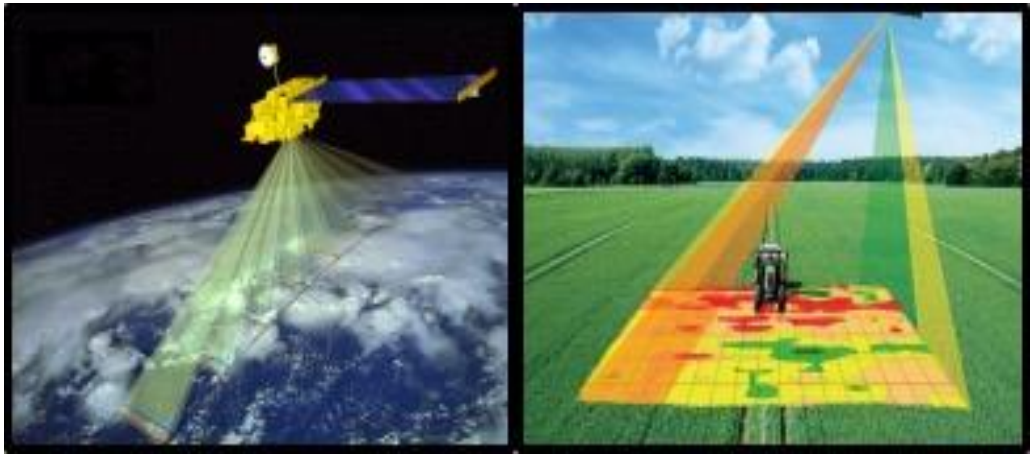


Precision Agriculture

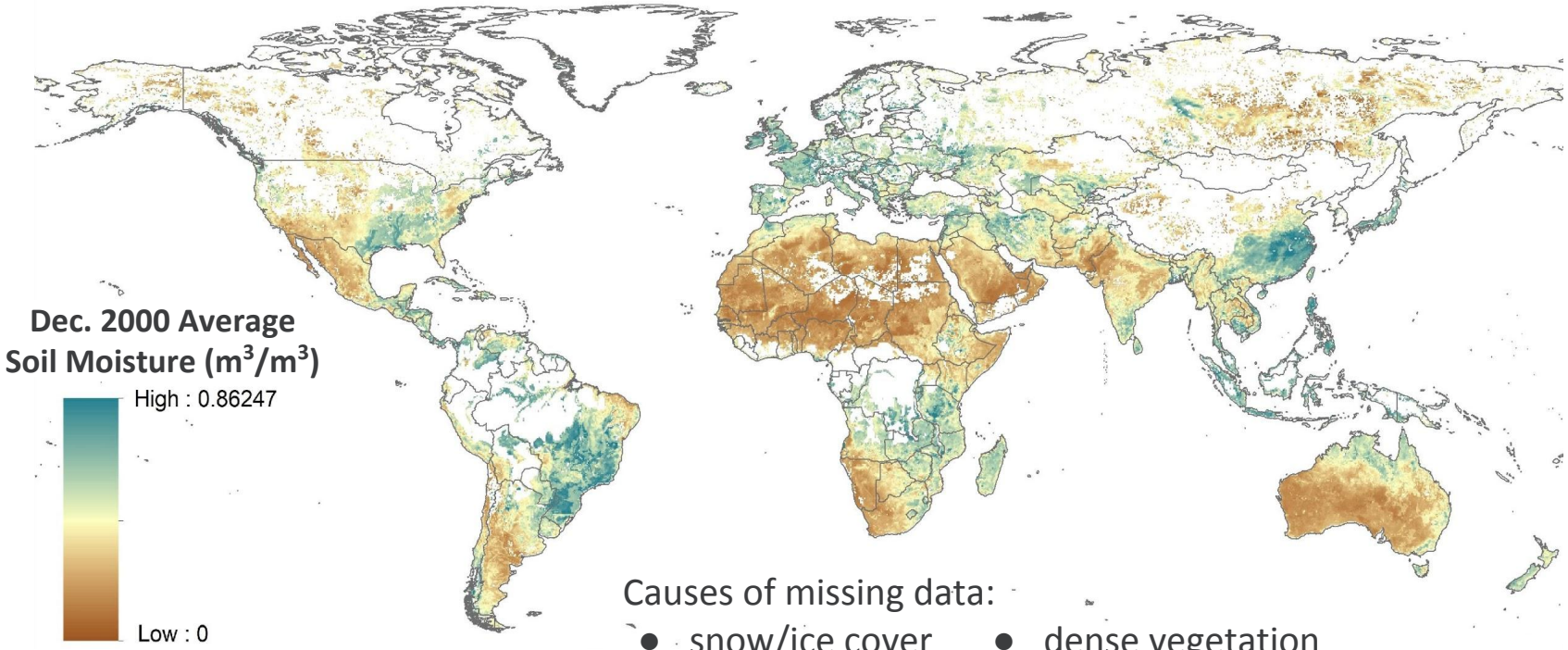
Image source: (left) GLOBE Carbon Cycle Project

Soil moisture data collection

- Satellite-borne remote sensing technology
  - Infrared to radio
  - Active and passive
- Data collected
  - Global coverage
  - Daily measurement



Challenge 1: incomplete soil moisture data

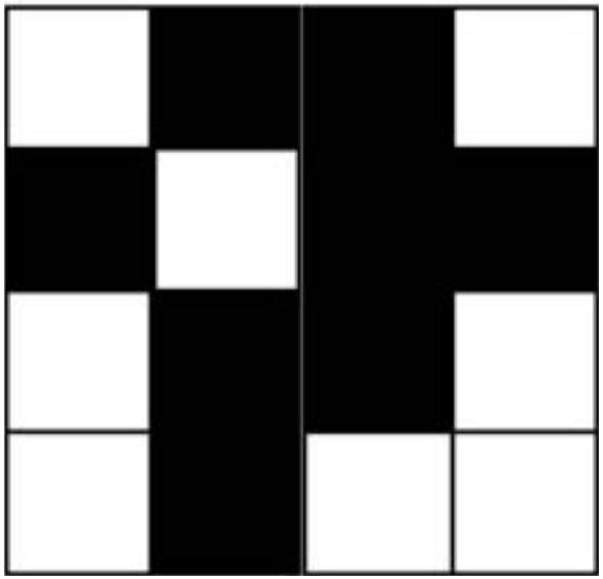


Causes of missing data:

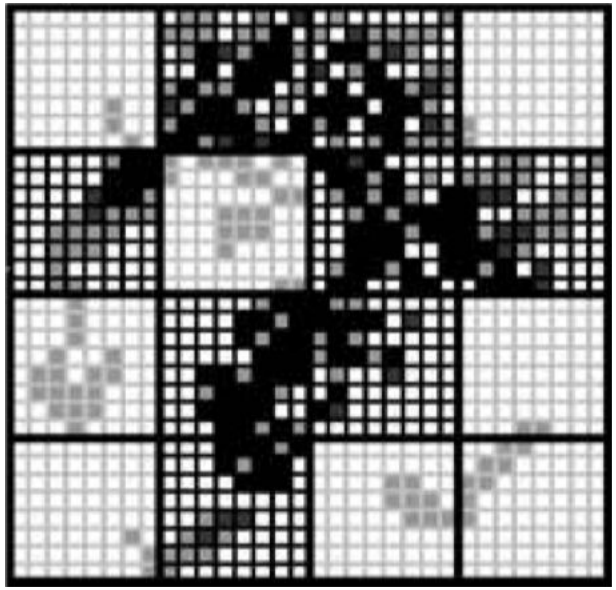
- snow/ice cover
- frozen surface
- dense vegetation
- extremely dry surface

Data source: ESA-CCI soil moisture database, http://www.esa-soilmoisture-cci.org

Challenge 2: coarse-grained soil moisture data



Original Resolution
27 km × 27 km

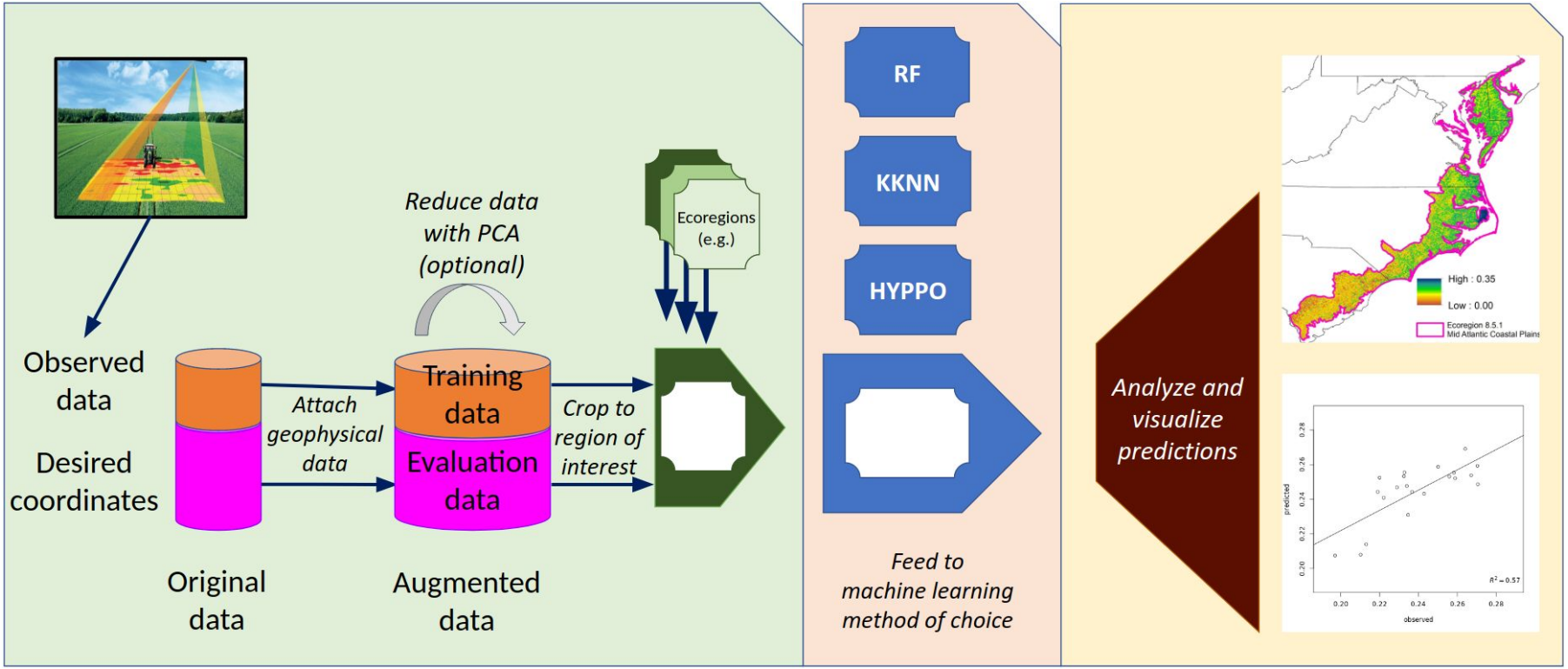


Desired Resolution
1 km × 1 km

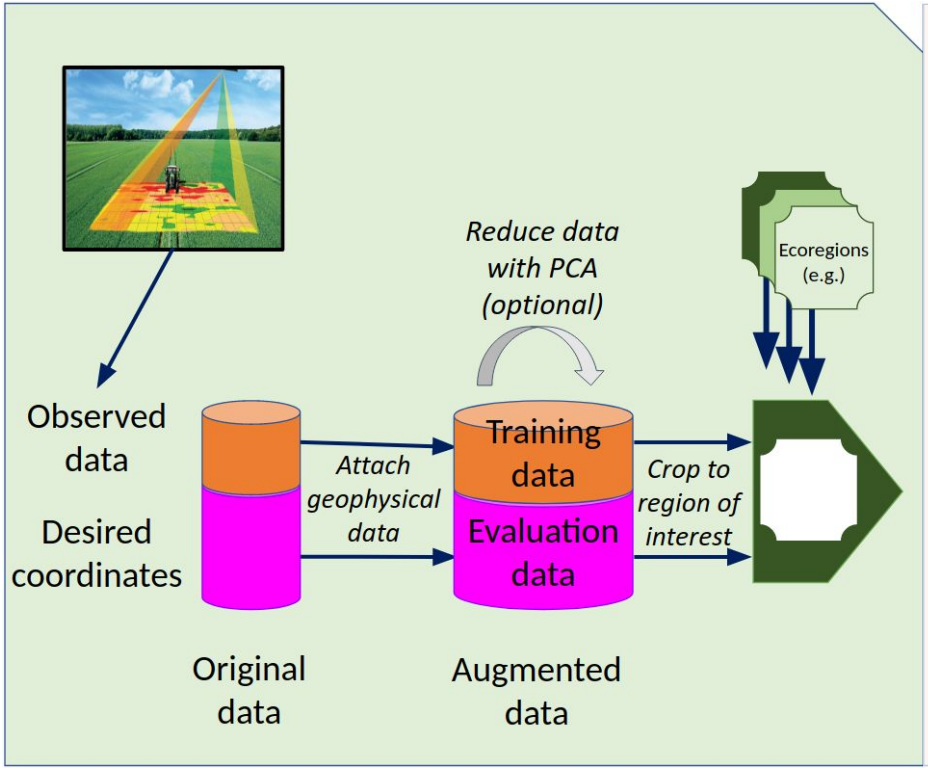
Image source: McPherson et al., *Using coarse-grained occurrence data to predict species distributions at finer spatial resolutions—possibilities and limitations*, Ecological Modeling **192**:499–522, 2006.

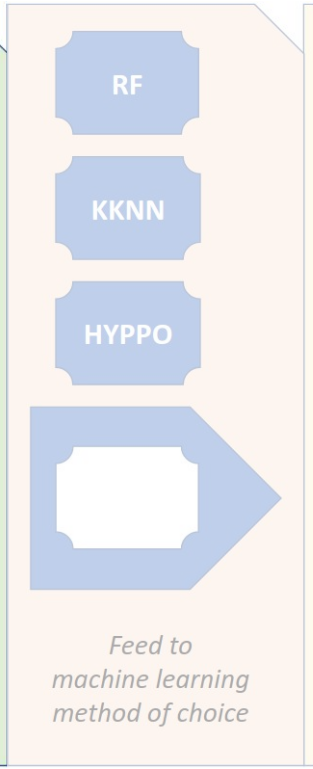
SOMOSPIE: a modular SOil MOisture SPatial Inference Engine

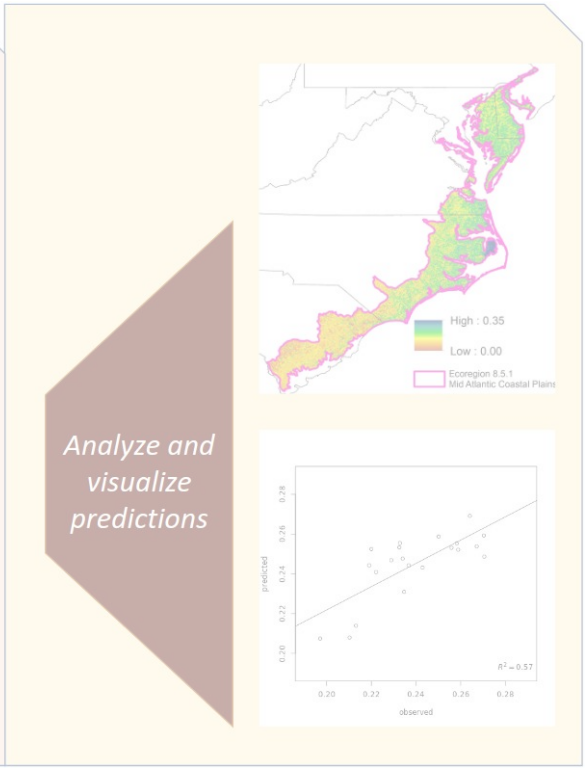
SOMOSPIE: SOil MOisture SPatial Inference Engine



SOMOSPIE: building from available data

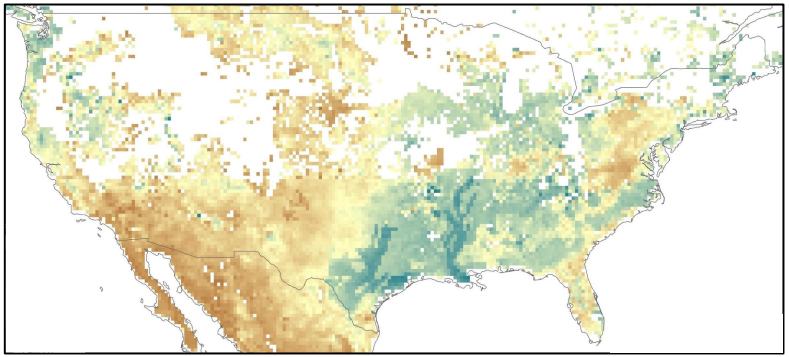






Input data

Coarse Resolution Soil Moisture Data



27 km x 27 km pixels

Each pixel is a 3-dimensional vector (x, y; sm):

- latitude and longitude of the centroid
- average soil moisture ratio in pixel

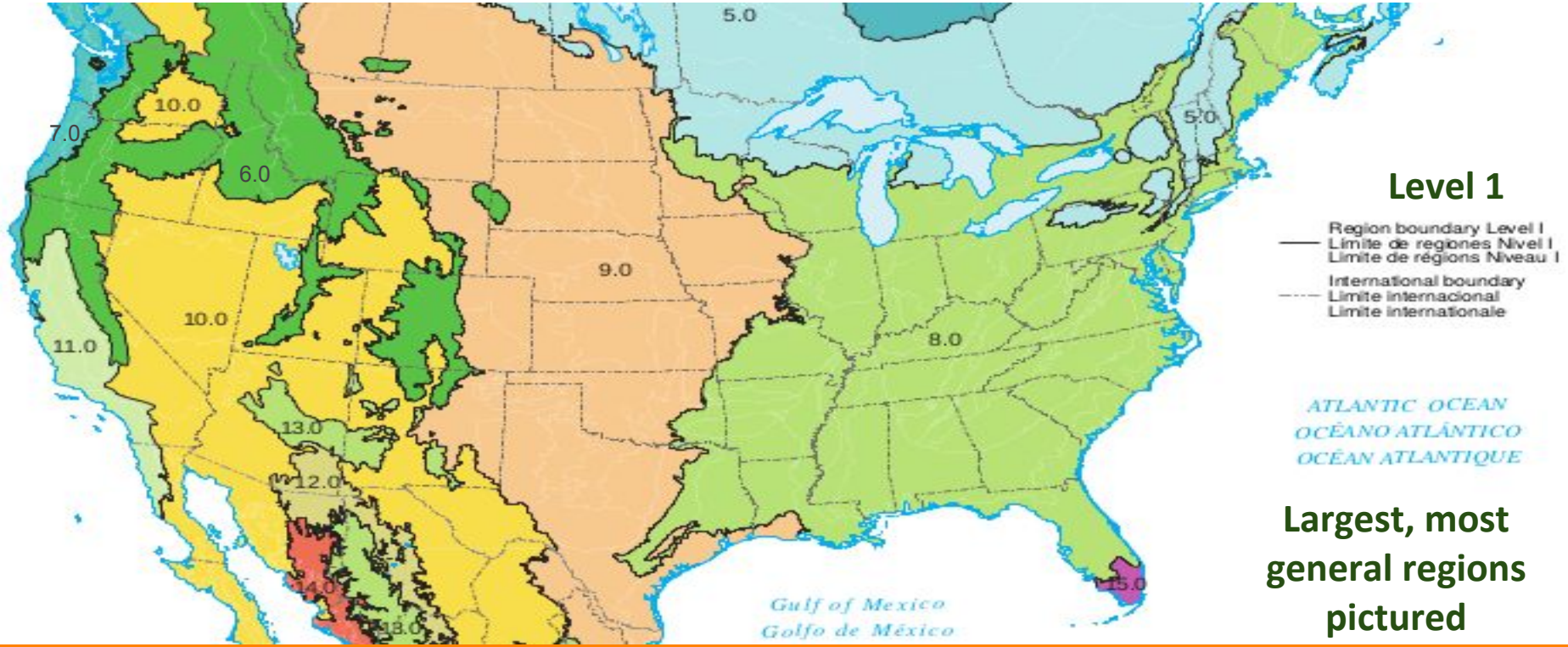
Fine Resolution Geophysical Data

1 km x 1 km pixels

Each pixel is a 17-dimensional vector (x, y; topos):

- latitude and longitude of the centroid
- 15 topographic parameters at the centroid

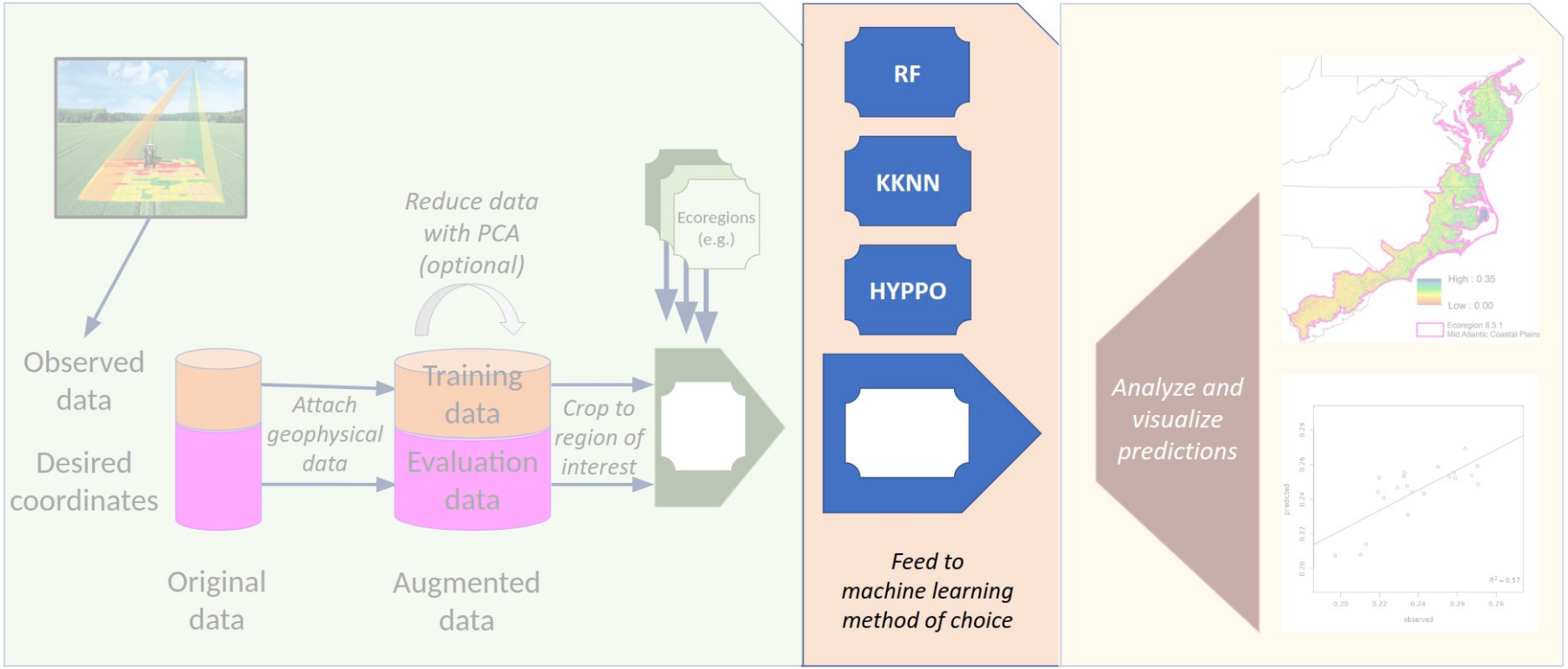
CEC ecoregions



Largest, most general regions pictured

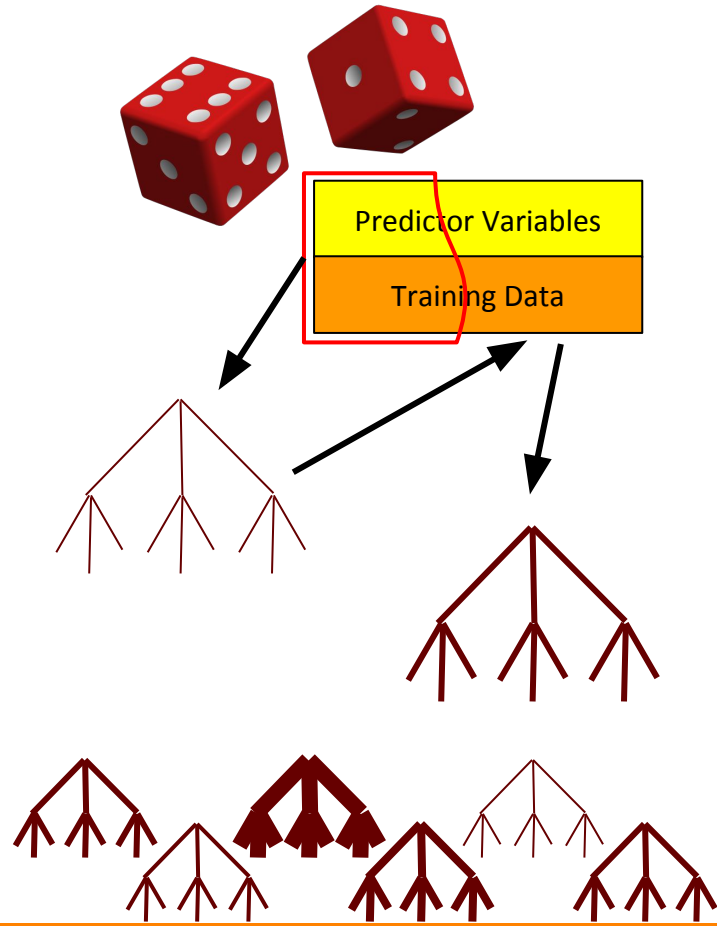
CEC: Commission for Environmental Cooperation
Image source: cec.org ecoregion documentation

SOMOSPIE: leveraging machine learning



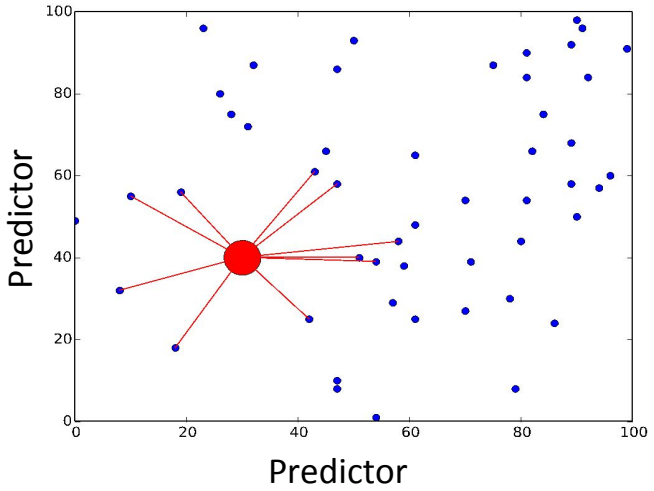
RF: Random Forests

- 500x:
  - Randomly select subset of predictors and training data
  - Grow a decision tree
  - Test the tree against the unused training data
  - Assign a weight to the tree based on its accuracy
- Compute weighted mean of all 500 predictions



KKNN: Kernel-weighted K-Nearest Neighbors

Select k nearest neighbors from training data



Traditional k-NN

- **Manually select k and distance metrics**
- **Compute mean of the neighbors' values**

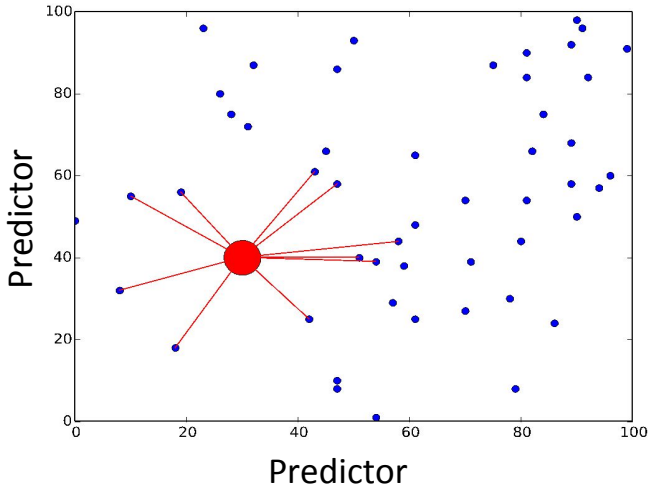
$$z(\vec{x}) := \frac{1}{k}\sum_{j=1}^{k} z_{i_j}$$

KKNN

- **Automatically select k and distance kernel using cross validation**
- **Compute weighted means with the kernel**

HYPPO: HYbrid Piecewise POlynomial

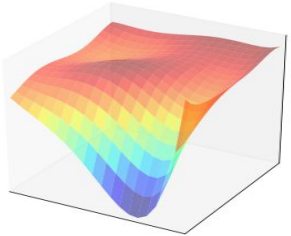
Select k nearest neighbors from training data



HYPPO

- Determine local polynomial degree using cross validation

- Use regression to generate a local polynomial model of that degree



Model comparison

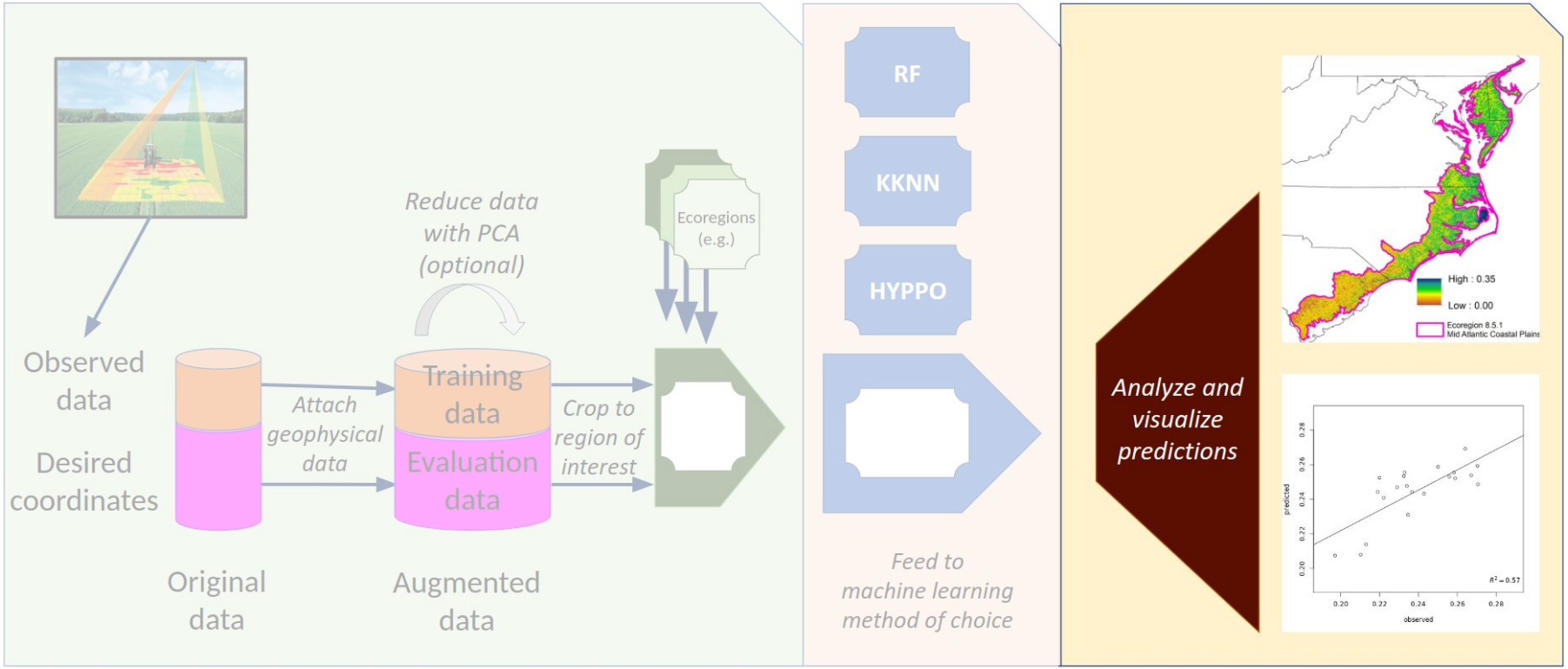
Parameters autoselected via 10-fold cross validation:

RF	KKNN	HYPPO
• number of predictor variables	• number of neighbors • weighting kernel	• polynomial degree

Fixed for entire modeling process

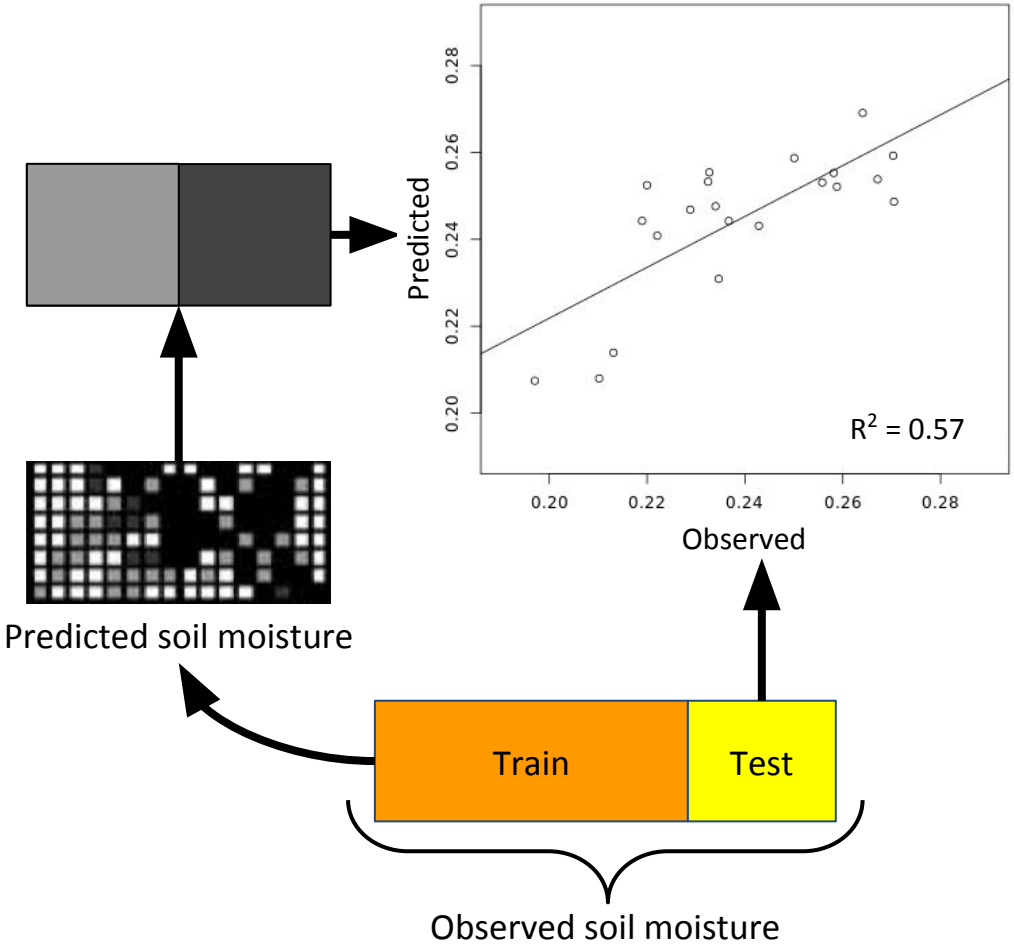
Selected for each local model

SOMOSPIE: analyzing predictions



Analysis

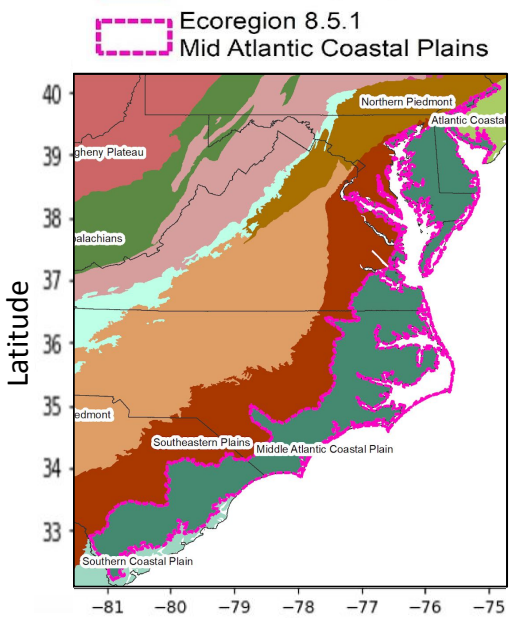
- Set aside 20% of the observed soil moisture for testing
- For each coarse pixel, compute the mean of all predicted values that fall within it
- Compute the correlation coefficient between the set of testing values and the means of the prediction values

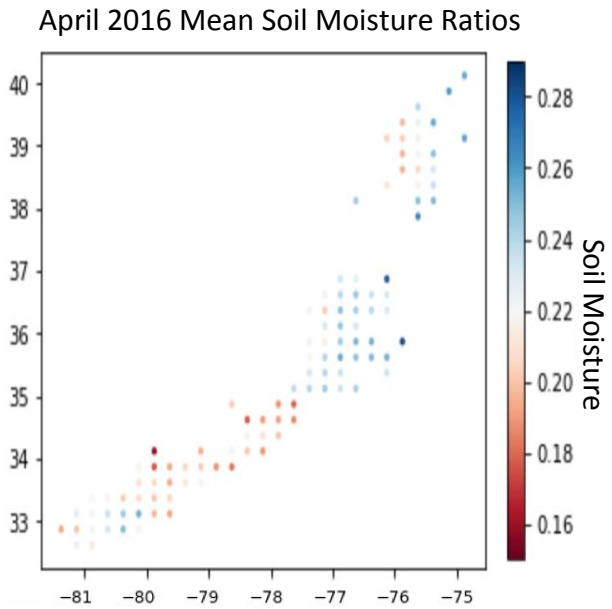


SOMOSPIE

Case study

Case study: one ecoregion, one month





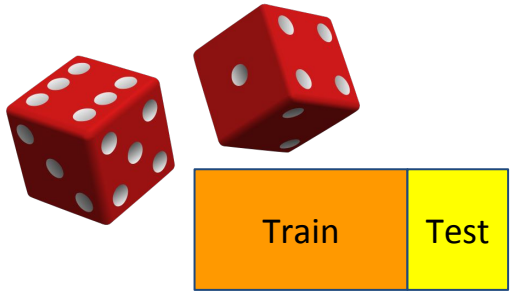
Research Questions

- How does climate affect soil moisture prediction?
- How does topography affect soil moisture prediction?

Methodology

10×

- Repeat each prediction 10 times
- Each iteration, randomly select 20% of data to be removed for testing
- Report the mean of the 10 r^2 values comparing the predictions to the testing data

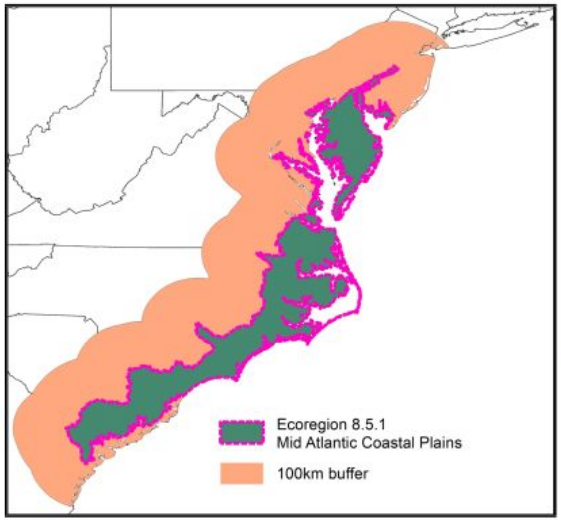


$$\frac{r_1^2 + r_2^2 + \cdots + r_{10}^2}{10}$$

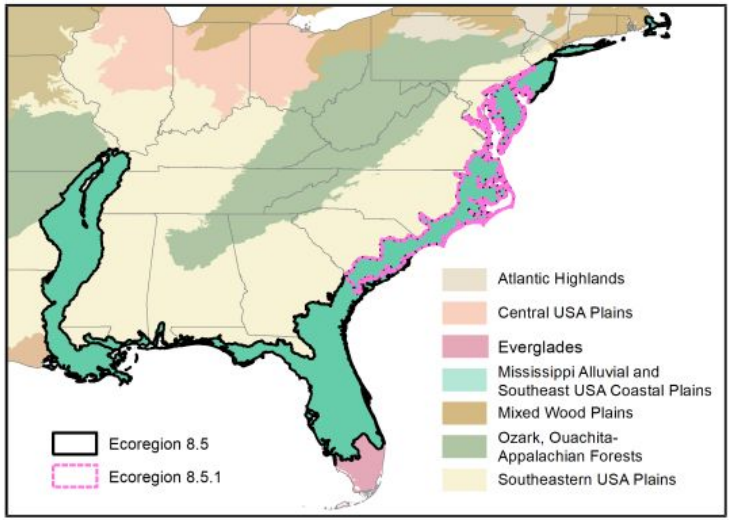
Case study: region

Adjusting the training region

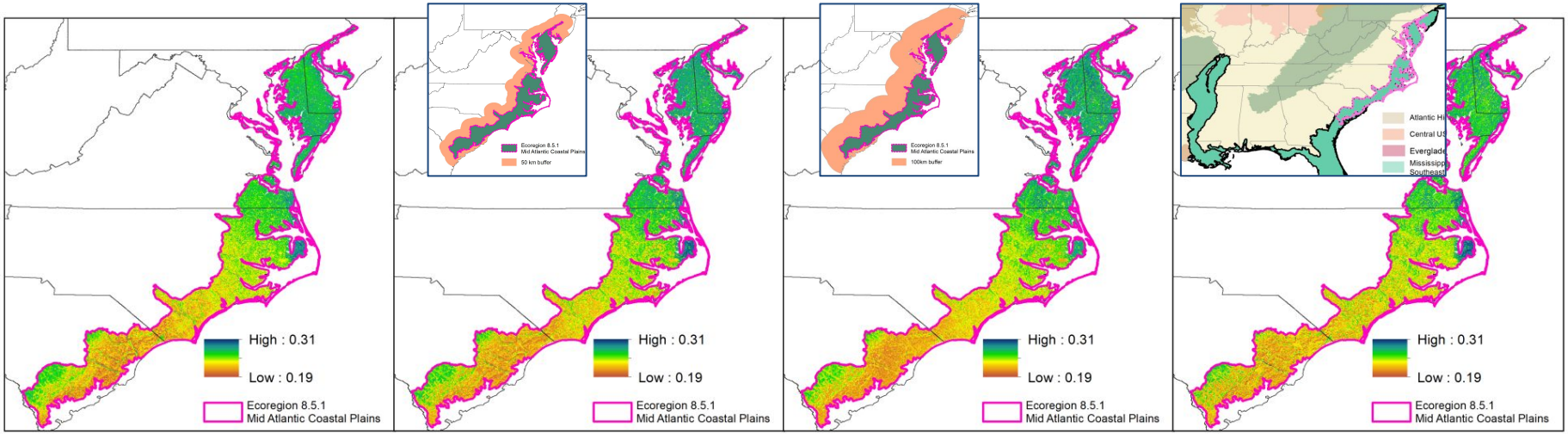
Add Fixed Buffer



Use More General Level



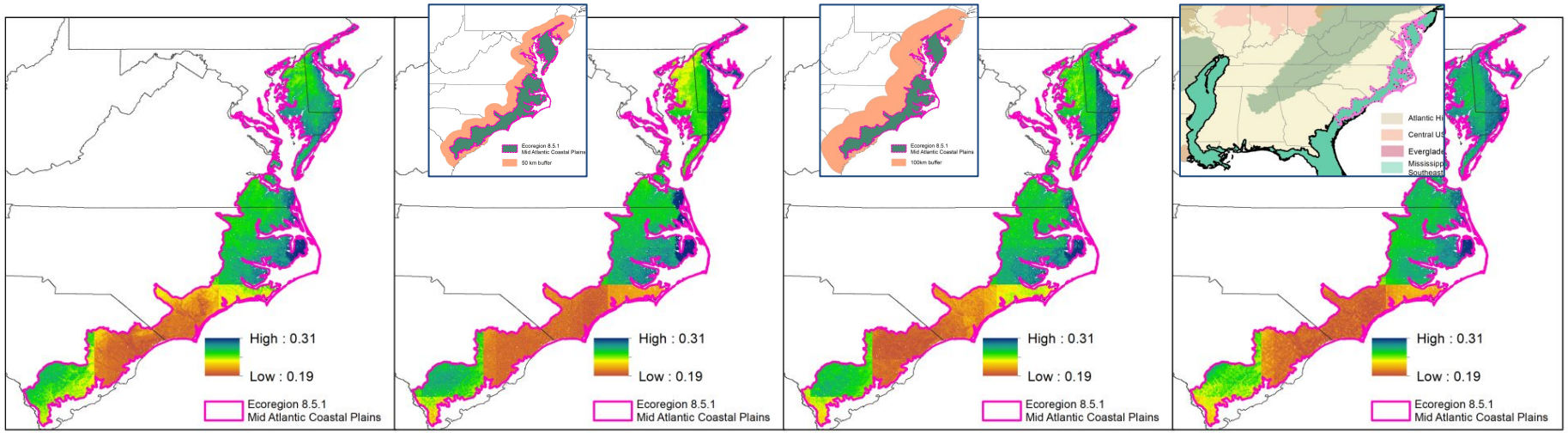
Adjusting the training region for KKNN



(a) Level III 8.5.1 (b) 8.5.1 with 50 km buffer (c) 8.5.1 with 100 km buffer (d) Level II 8.5

mean R^2 = 0.30 mean R^2 = 0.24 mean R^2 = 0.20 mean R^2 = 0.10

Adjusting the training region for RF



(a) Level III 8.5.1 (b) 8.5.1 with 50 km buffer (c) 8.5.1 with 100 km buffer (d) Level II 8.5

mean R^2 = 0.58 mean R^2 = 0.58 mean R^2 = 0.60 mean R^2 = 0.55

Case study: topography

Topographic Data Reprise

Fine Resolution Geophysical Data

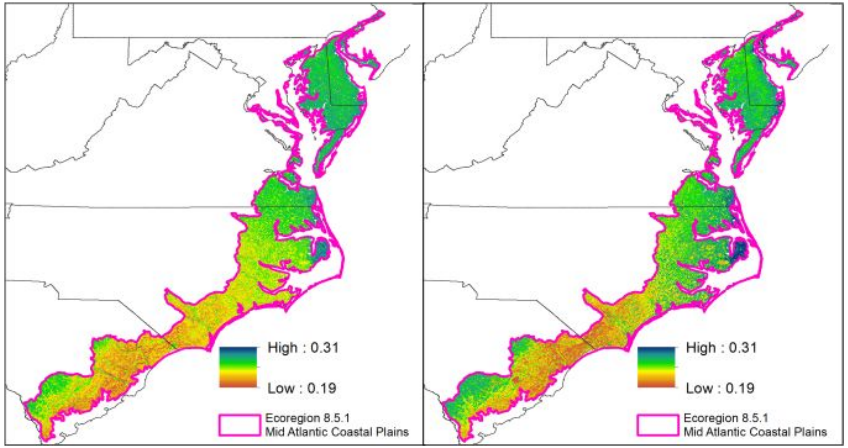
1 km x 1 km pixels

Each pixel is a 17-dimensional vector (x, y; topos):

- latitude and longitude of the centroid
- 15 topographic parameters at the centroid

Use PCA to reduce data redundancy (15 → 6 or 7)

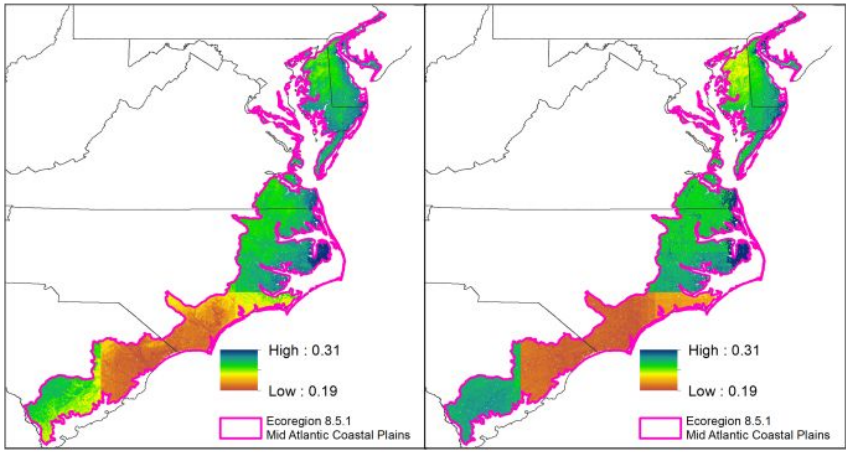
KKNN



without PCA | with PCA

mean R^2 = 0.30 | mean R^2 = 0.29

RF



without PCA | with PCA

mean R^2 = 0.58 | mean R^2 = 0.69

Case study: digging deeper

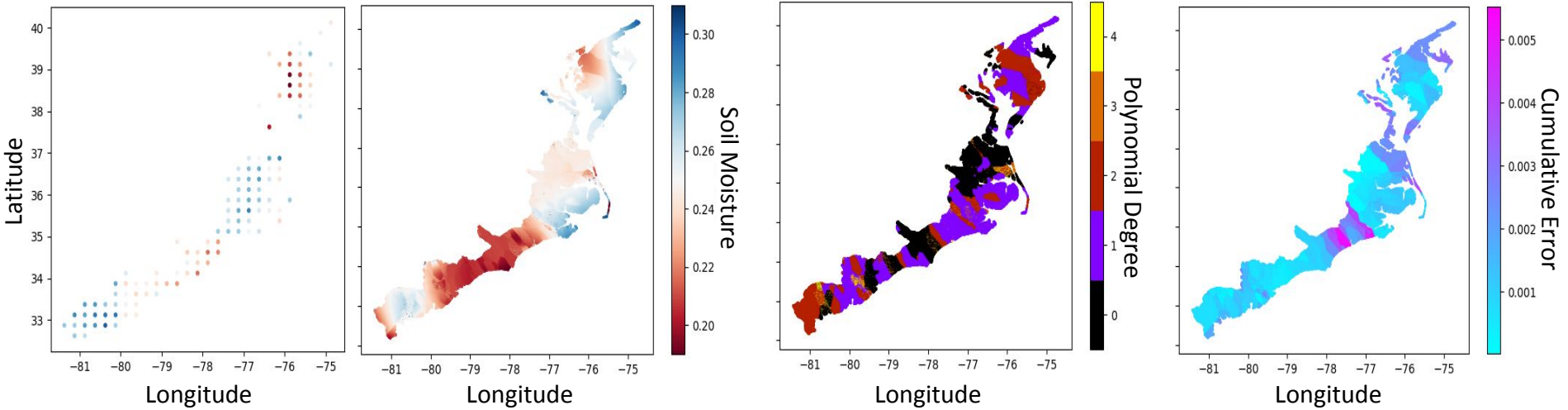
Uncovering Model Features with HYPPO

April 2016 mean soil moisture ratios

Predicted Soil Moisture Heatmap

Degrees of local polynomial models (selected by 10-fold cross validation)

Minimum cumulative SSE during 10-fold cross validation



Big picture

Lessons Learned

Region effect

- Selection of testing region can significantly influence prediction, as demonstrated by the decreased accuracy of KKNN as the testing region grew; this affirms climate influence

Topography effect

- Data redundancy can be reduced via PCA without greatly affecting KKNN prediction accuracy, and may dampen overfitting by RF