Lecture 9:
Modeling Data with KNN

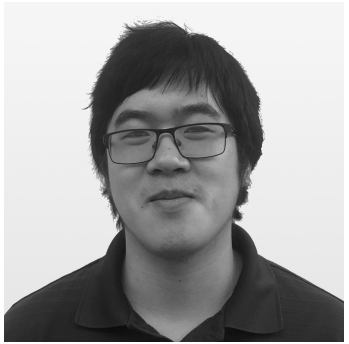
COSC 526: Introduction to Data Mining
Spring 2020

Instructor:

Michela Taufer

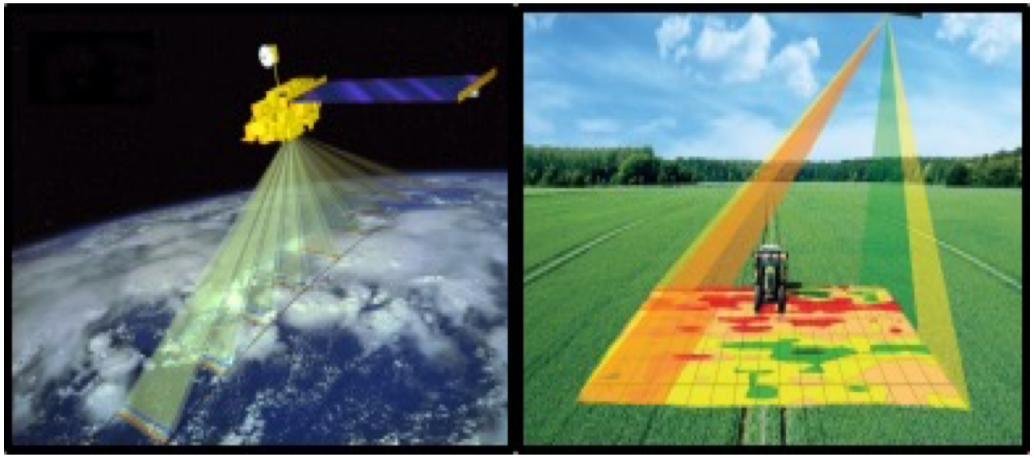
GRA:

Mike Wyatt



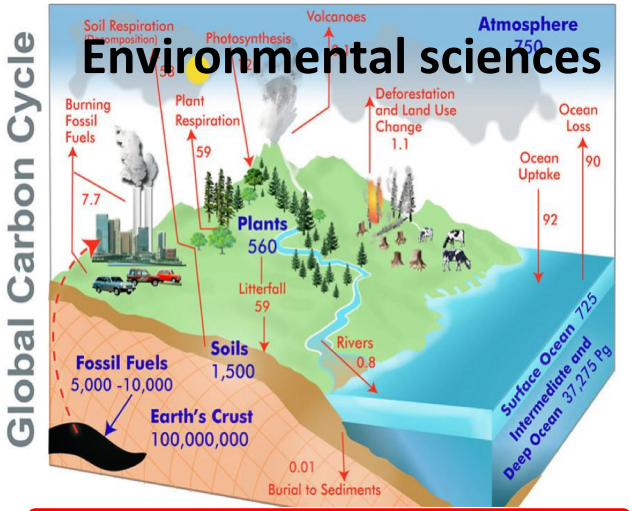
Nigel Tan

Relevance of soil moisture data



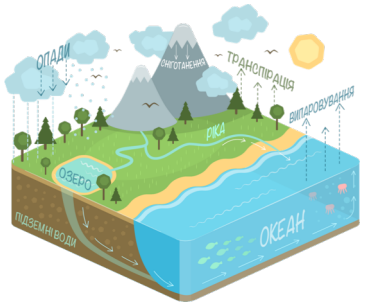
- Satellite-borne remote sensing technology
  - Infrared to radio
  - Active and passive

Environmental sciences



Workflows for precision agriculture

Application



Soil moisture leveraged for:
- Environmental sciences
- Precision agriculture

Computation

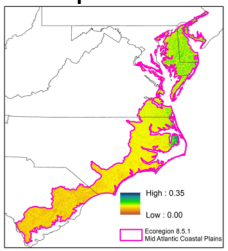
Data Feedback

Fine-grained Soil Moisture
Landscape Surface DSM
Weather Data NOAA

Fine-grained, complete data

Analytics representations + algorithms

Data prediction

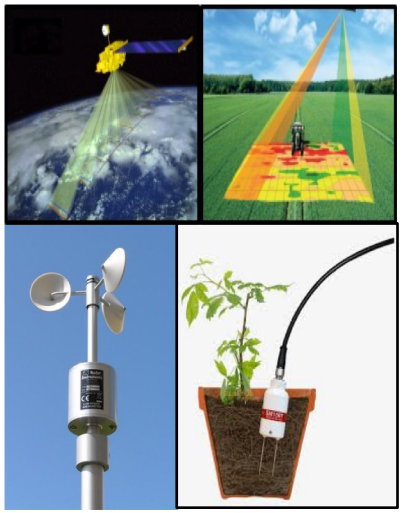


Data Analytics

Soil Moisture ESA-CCI
Landscape Surface DSM
Weather Data NOAA

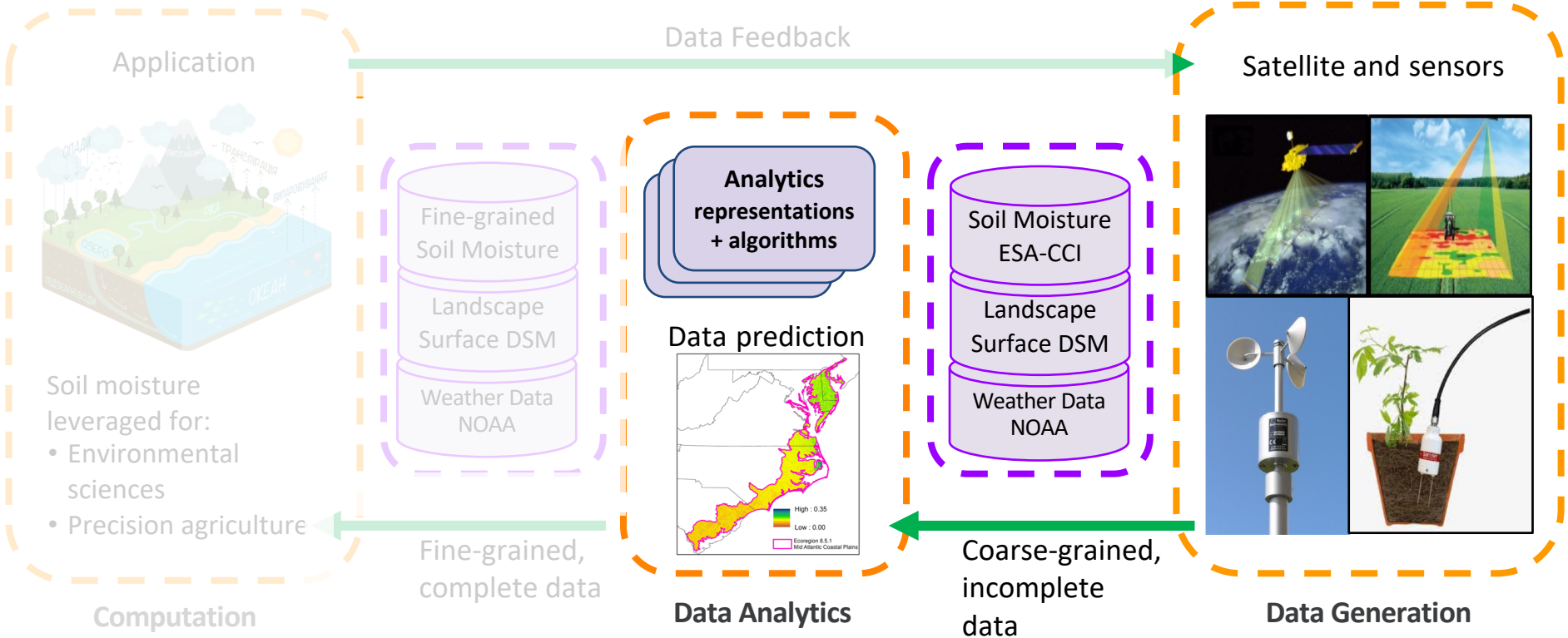
Coarse-grained, incomplete data

Multiscale data

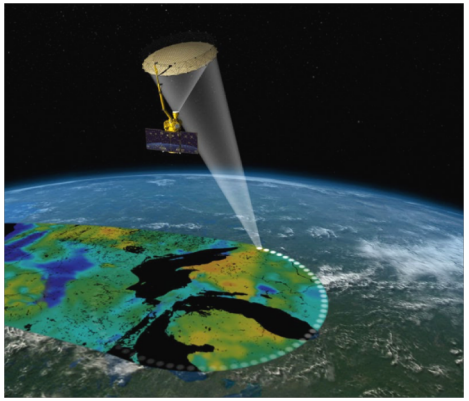


Data Generation

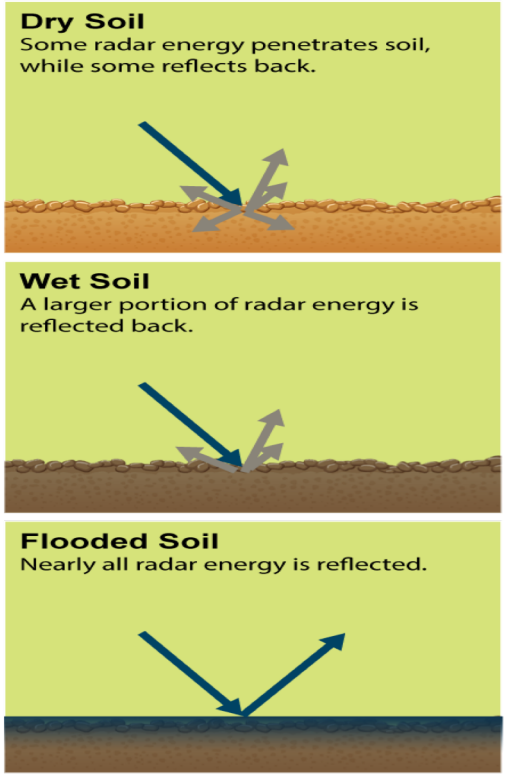
Data analytics for soil moisture

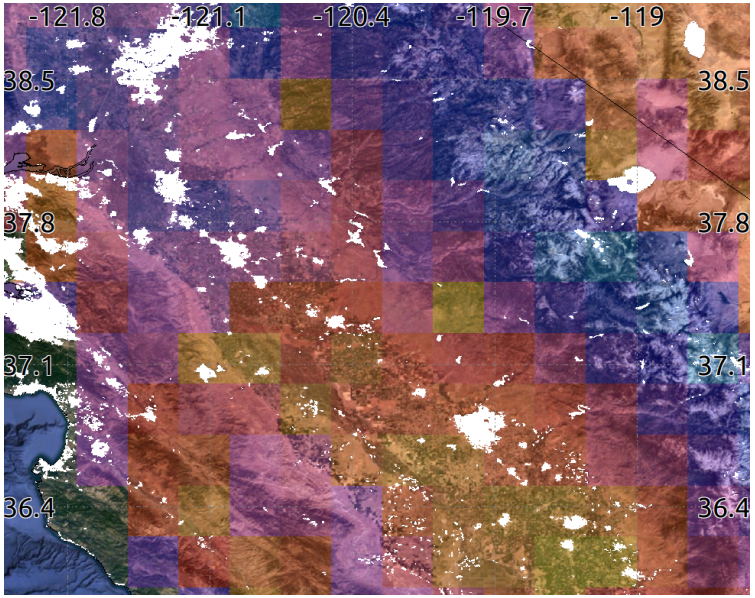


Challenge 1: incomplete soil moisture data (I)



Satellites collect raster data across the surface of the Earth

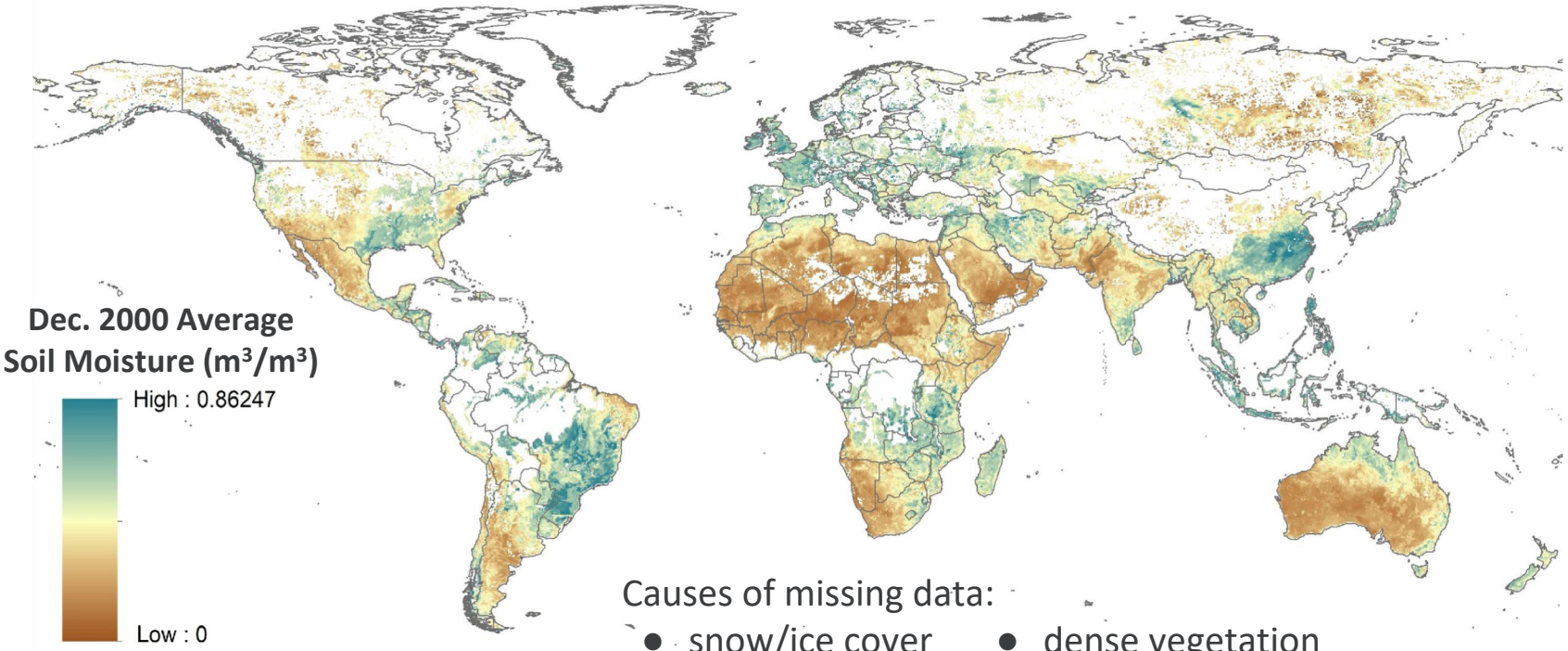




Visualization example of the ESA-Climate Change Initiative Soil Moisture database with a coarse pixel size of 27x27km

(Liu et al. 2011 HESS, Liu et al. 2012 RSE)

Challenge 1: incomplete soil moisture data (II)

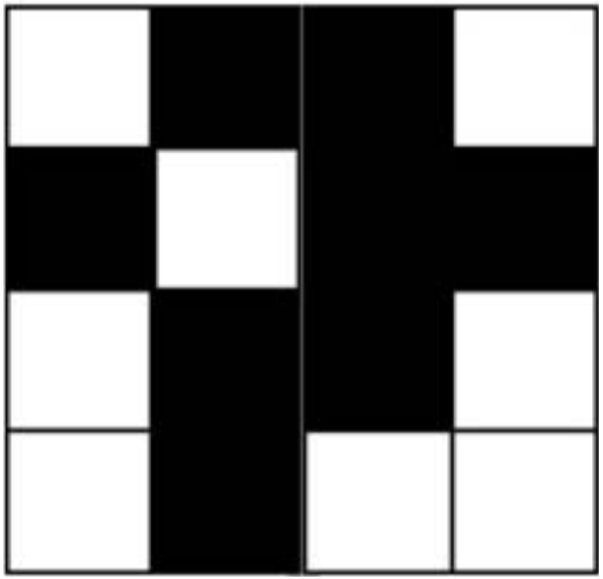


Causes of missing data:

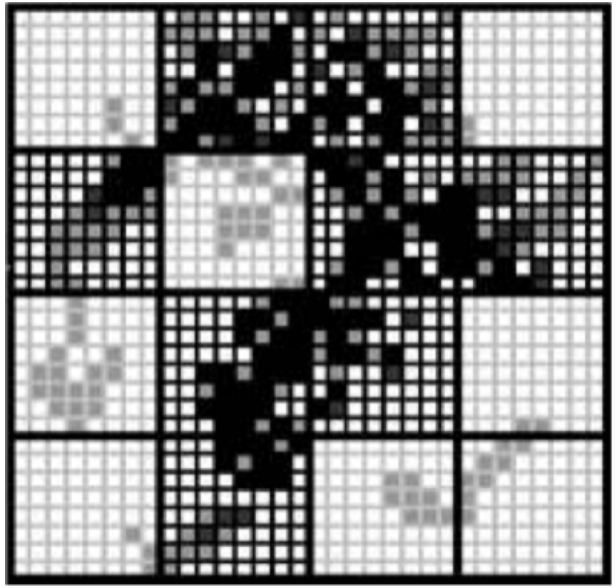
- snow/ice cover
- frozen surface
- dense vegetation
- extremely dry surface

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

Challenge 2: coarse-grained soil moisture data (I)



Original Resolution
27 km × 27 km

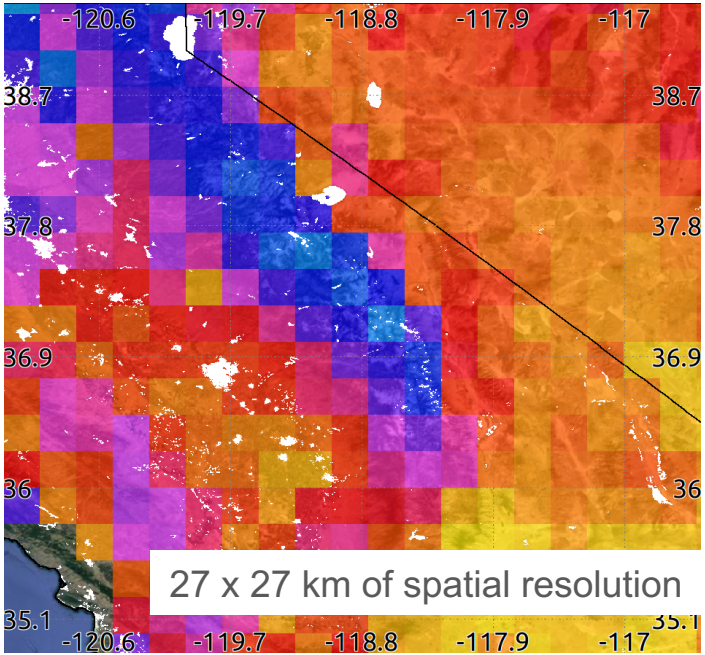


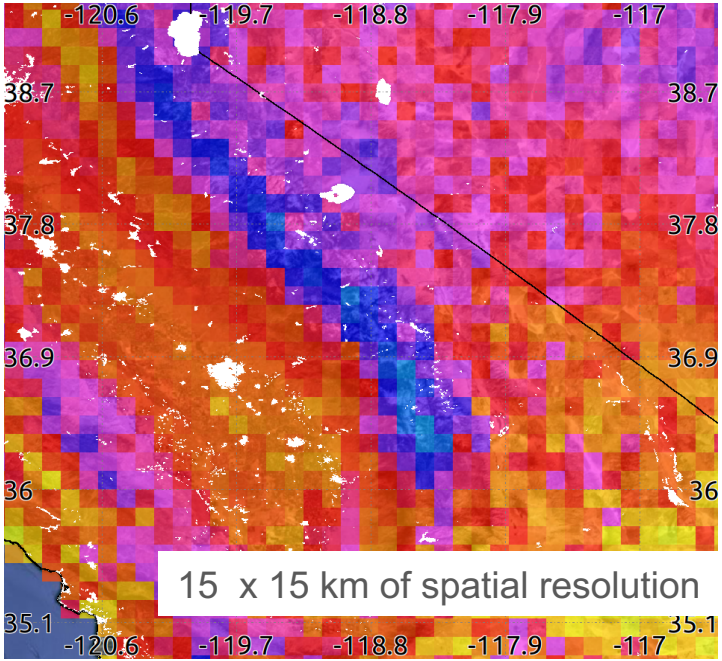
Desired Resolution
1 km × 1 km

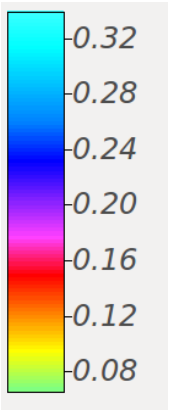
Image source: McPherson et al., *Using coarse-grained occurrence data to predict species distributions at finer spatial resolutions—possibilities and limitations*, Ecological Modeling **192**:499–522, 2006.

Challenge 2: coarse-grained soil moisture data (II)

Original product ESA CCI (m^3 m^{-3}, mean 2013)

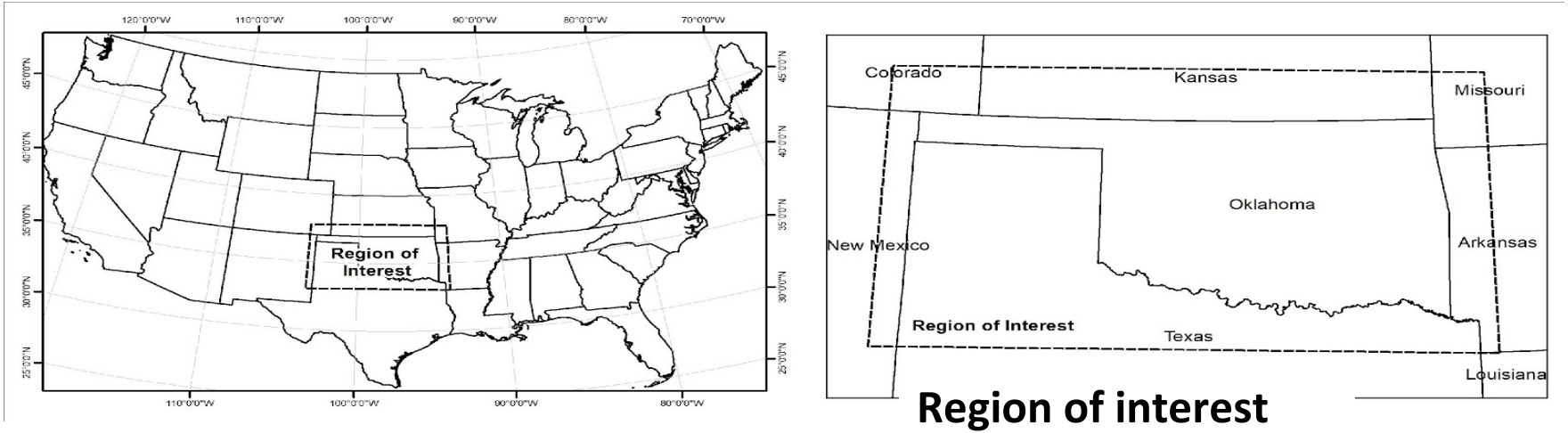






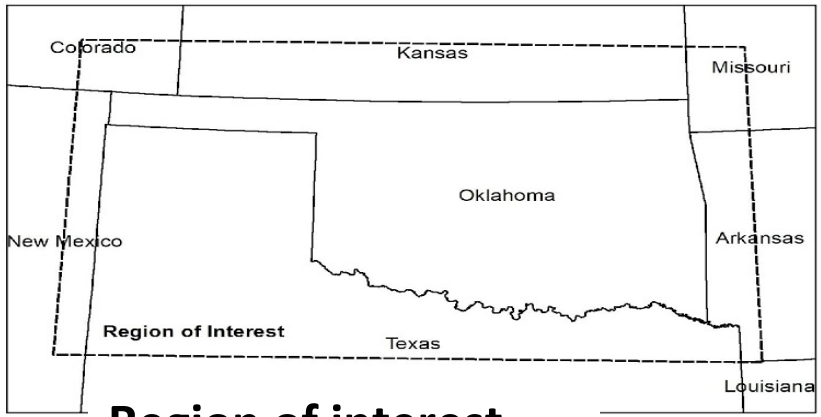
M. Guevara , M. Taufer, and R. Vargas. Gap-Free Annual Soil Moisture Global across 15km Grids: 1991-2016. Earth System Science Data, 2019.

Selecting a region of interest

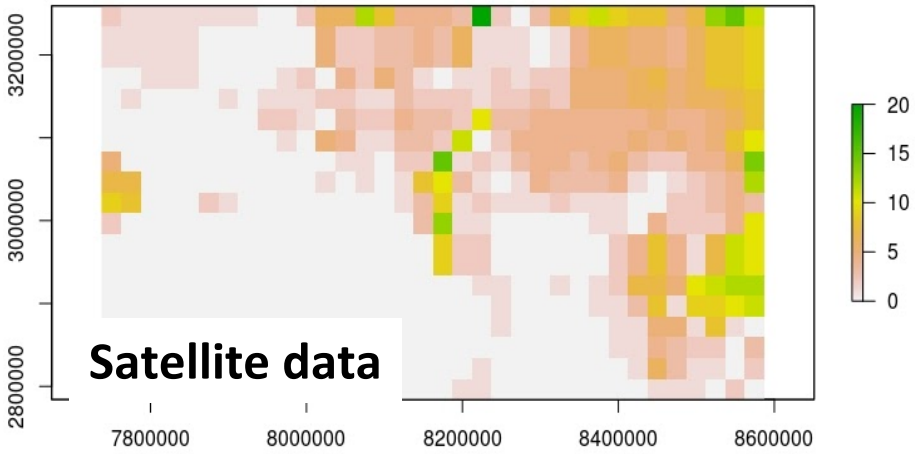


Region of interest

Integration of multiscale data: from satellites …



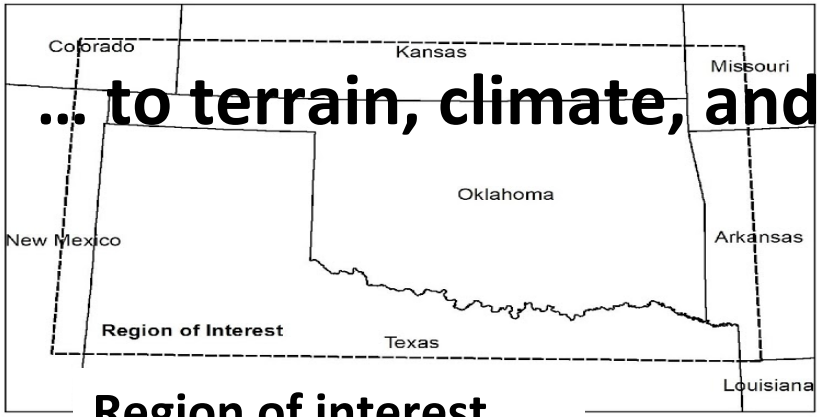
Region of interest



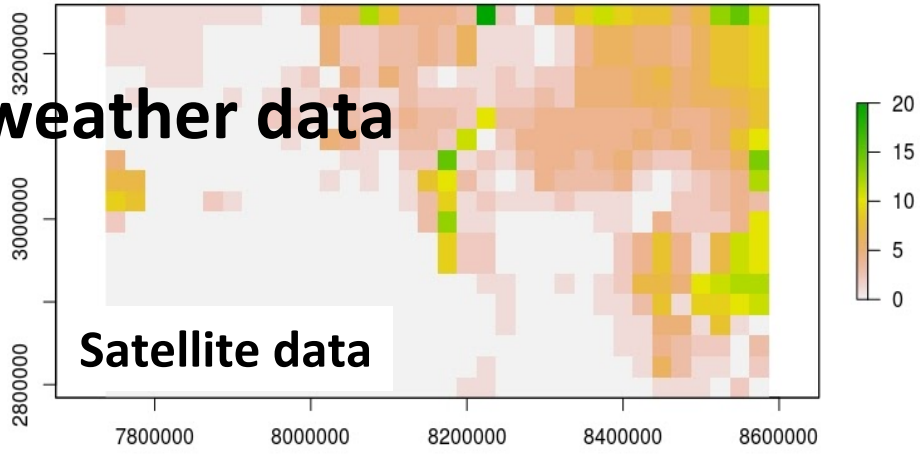
Satellite data

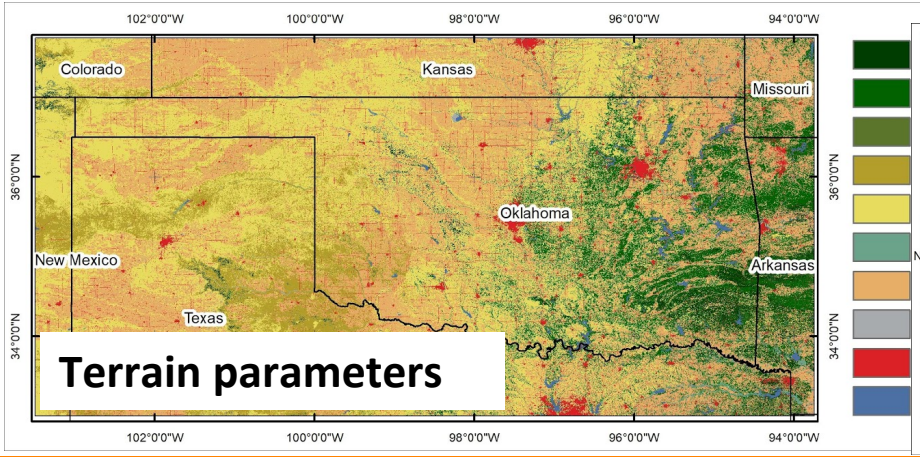
R Llamas, M Guevara, D Rorabaugh, M Taufer, R Vargas. Spatial Gap-Filling of ESA CCI Satellite-Derived Soil Moisture based on Geostatistical Techniques and Multiple Regression. Remote Sensing, 2020.

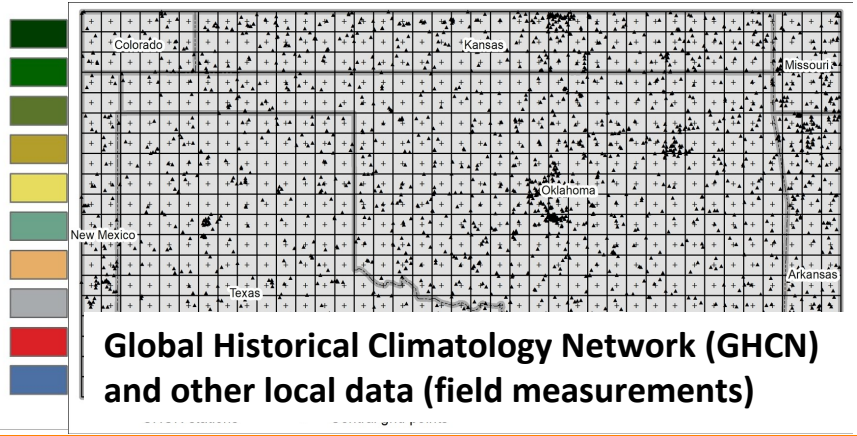
… to terrain, climate, and weather data



Region of interest

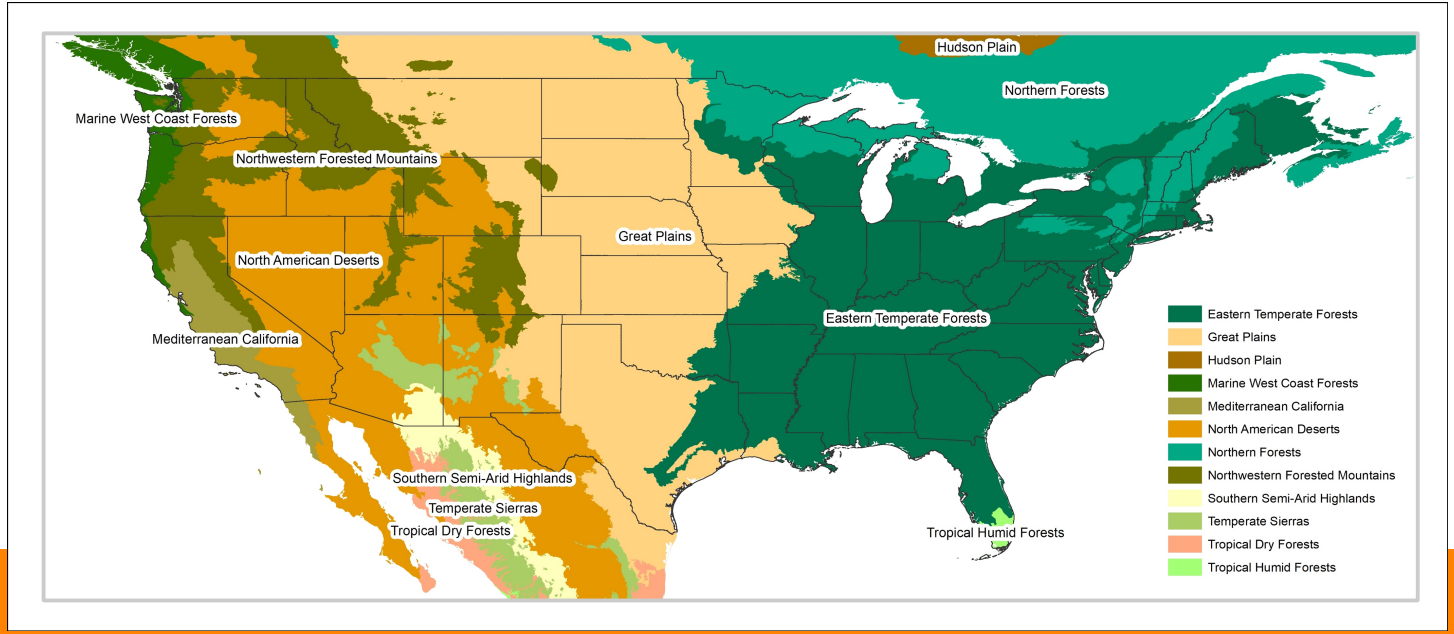






Integrate Climate Regions in Workflow

- Ecoregions across the conterminous United States play a key role to simplify the prediction process



Dealing with Heterogenous Data

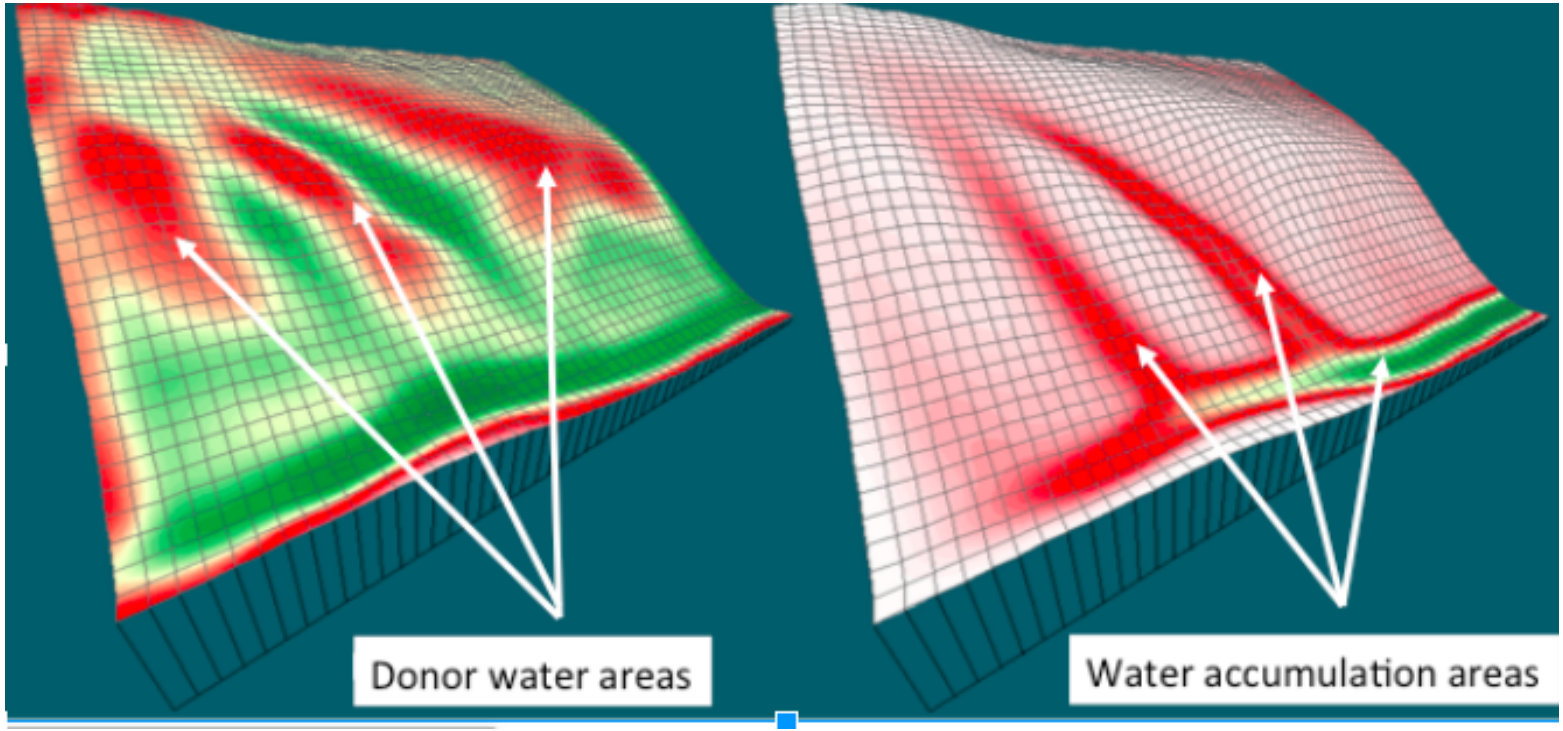
- Satellite date is from European Space Agency
- Land surface characteristic (topography data) are local features
- Ecoregion boundary allows the simplification of predictions
  - When working inside an ecoregion we can ignore the climate impact and heavily rely on topographical data

Dataset	Spatial resolut.	Temporal resolution	Variable / Description	Source
ESA-CCI	0.25 degrees	Daily, 1978-2013	soil moisture (m^3/m^3)	European Space Agency
Digital surface model (DSM)	≈30 meters	Static ('Current')	Land surface characteristics	The Japan Aerospace Exploration Agency
CEC	n/a	Static ('Current)	Ecoregion boundaries	Commission for Environmental Cooperation

Dealing with Heterogenous Data

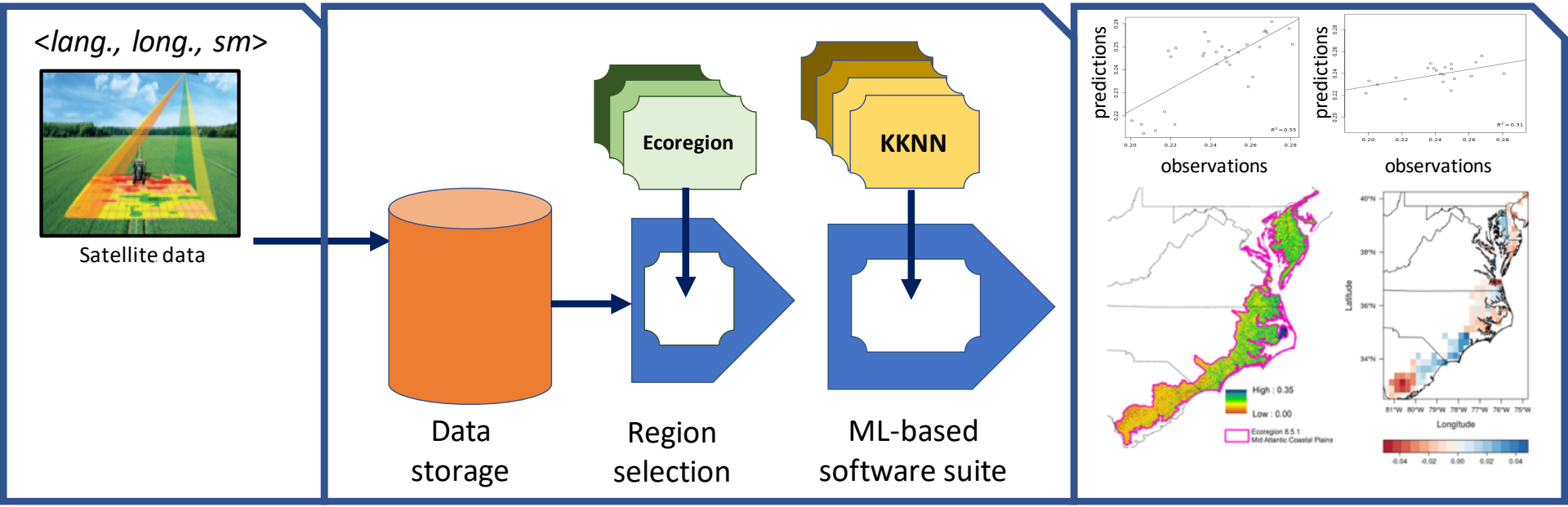
- Soil moisture data are collected by ESA satellites as part of its Climate Change Initiative
- Topographical parameters are derived from a Digital Elevation Model in SAGA GIS
  - Surrogate of overland flow of water (e.g., terrain slope)
  - Surrogate of potential incoming solar radiation (e.g., south or north slopes)

Example of terrain parameters: water wetness index



Shaw et al., 2016 GRL, Moore 2012, Geomorphology.

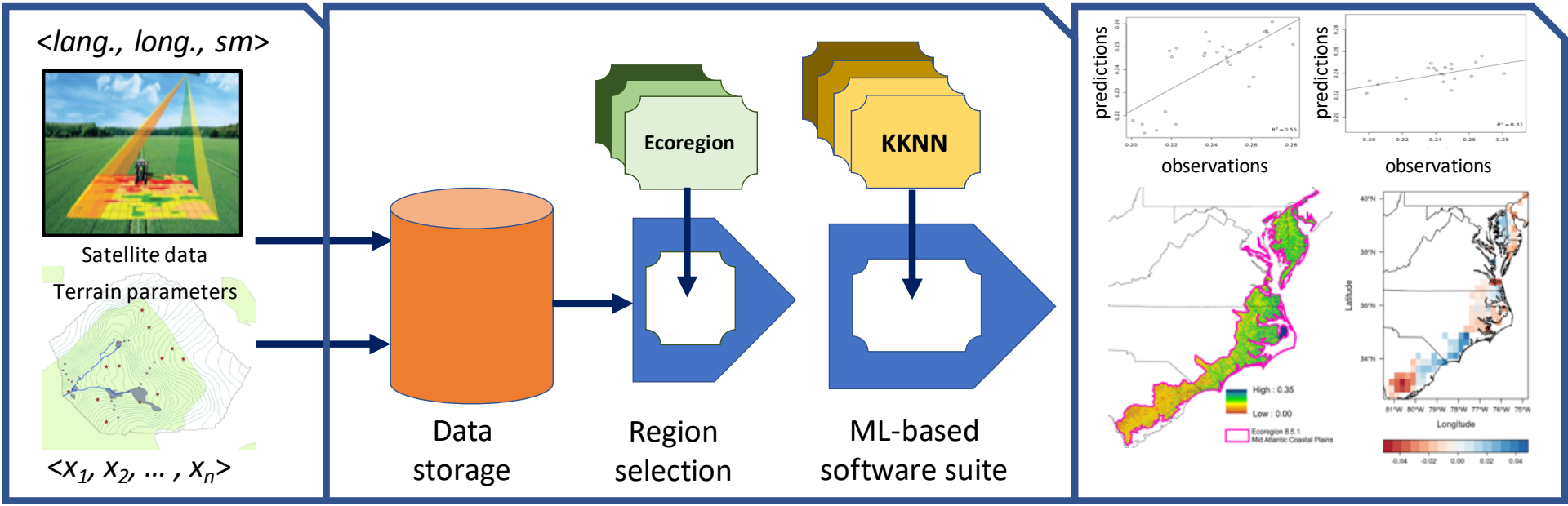
SOMOSPIE: SOil MOisture SPatial Inference Engine



Data collection **Feature extraction** **Predictions** **Analysis tools**

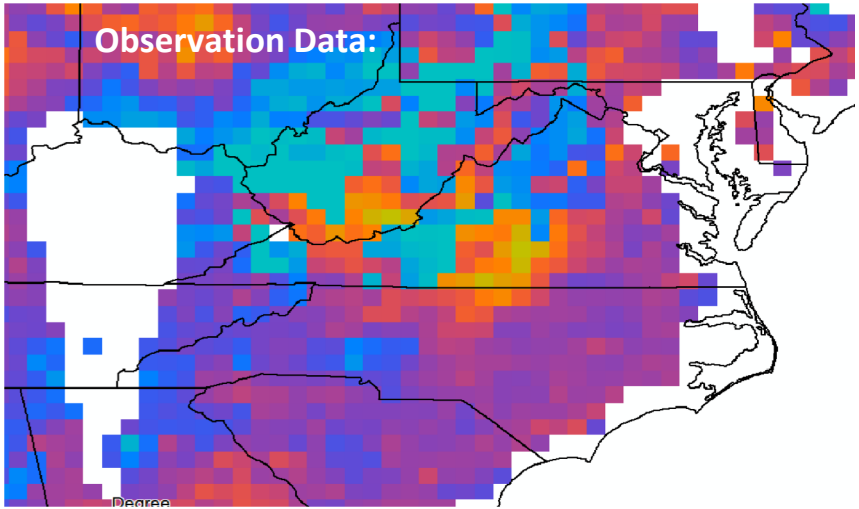
D. Rorabaugh, M. Guevara, R. Llamas, J. Kitson, R. Vargas, and **M. Taufer**. SOMOSPIE: A Modular SOil MOisture SPatial Inference Engine based on Data Driven Decisions. eScinece 2019.

SOMOSPIE: SOil MOisture SPatial Inference Engine



Data collection **Feature extraction** **Predictions** **Analysis tools**

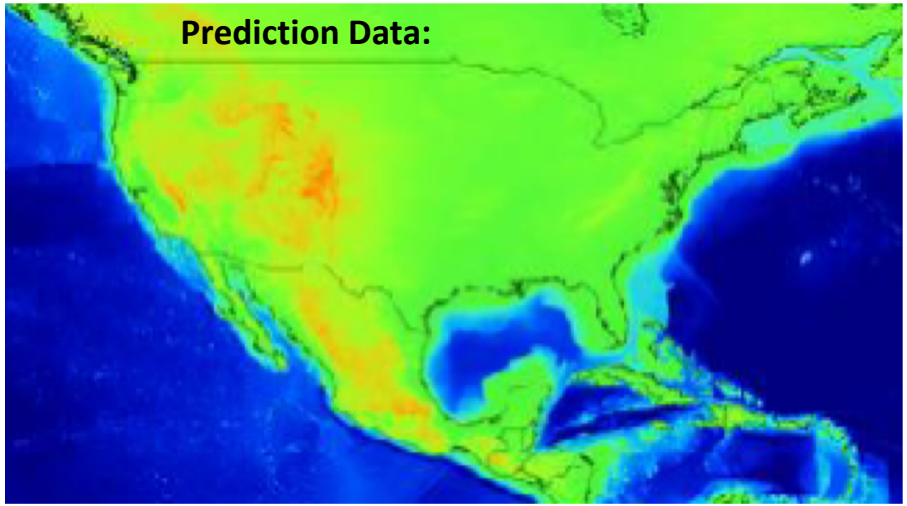
From coarse-grained observations to fine-grained predictions



27 km x 27km pixels

Each pixel is a vector:

- latitude and longitude of the centroid
- observed average soil moisture ratio in pixel
- 15 topographic parameters in centroid



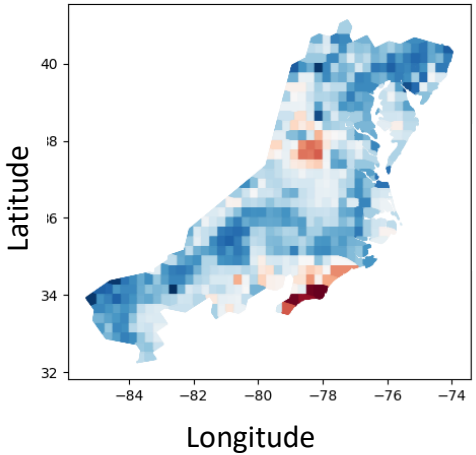
1 km x 1 km pixels

Each pixel is a vector:

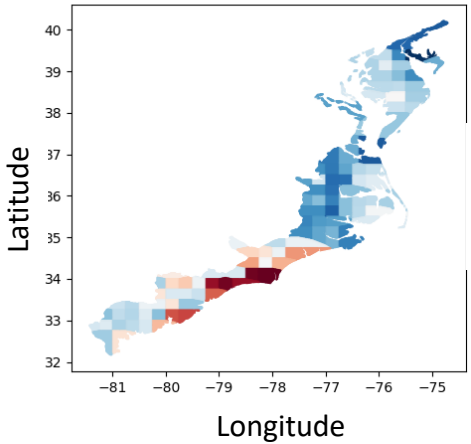
- latitude and longitude of the centroid
- predicted average soil moisture ratio in pixel
- 15 topographic parameters in centroid

Region selection: format of regions of interest

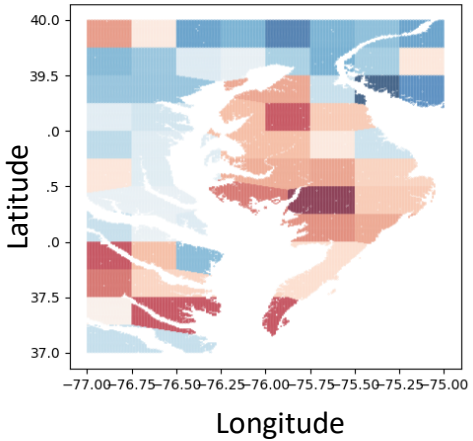
("NEON", "Mid Atlantic")



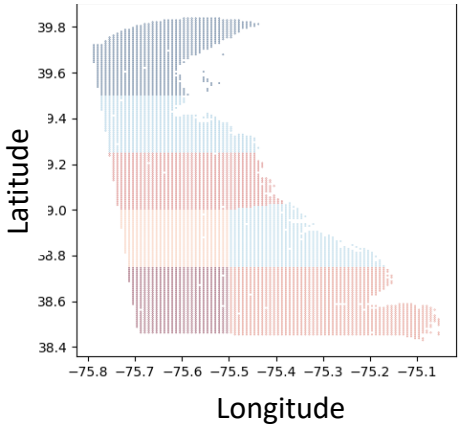
("CEC", "8.5.1")



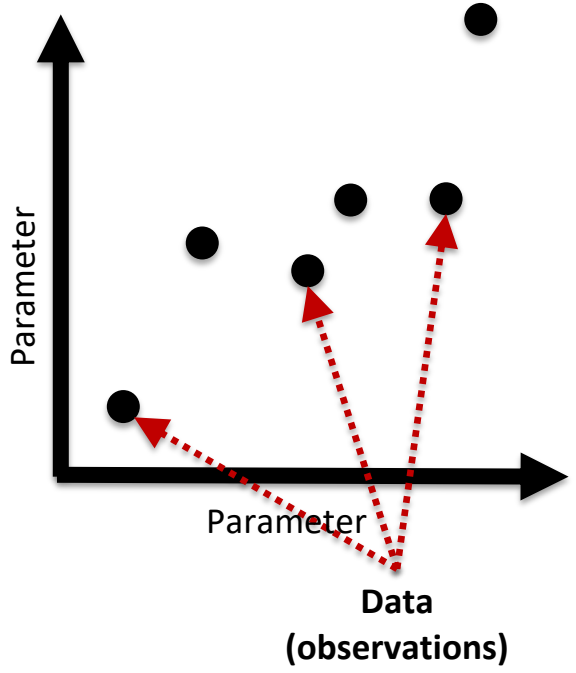
("BOX", "-77_-75_37_40")



("STATE", "Delaware")

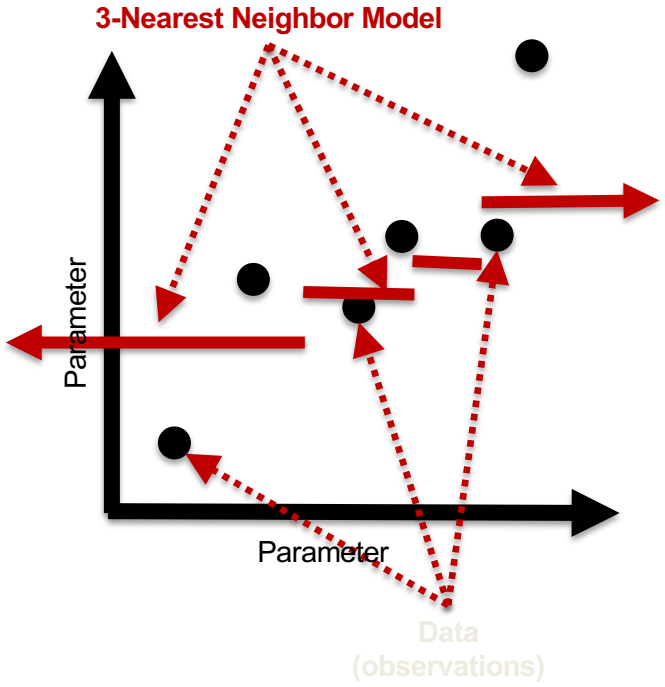


K Nearest Neighbors Model



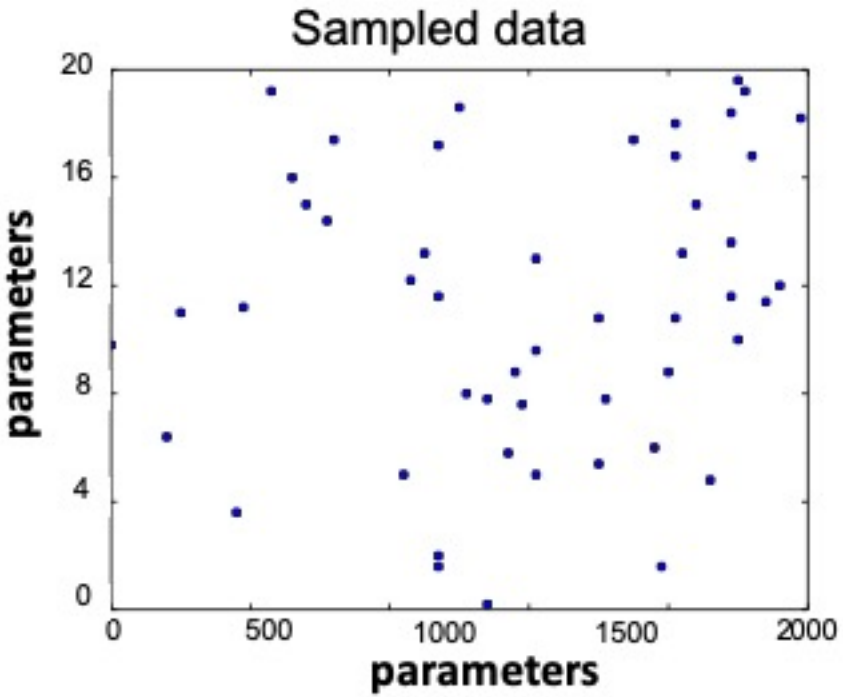
- Data → small amount of data to create simple, **local models**
- Model → **average** of k nearest neighbors
- Best for → surfaces that are locally **flat**

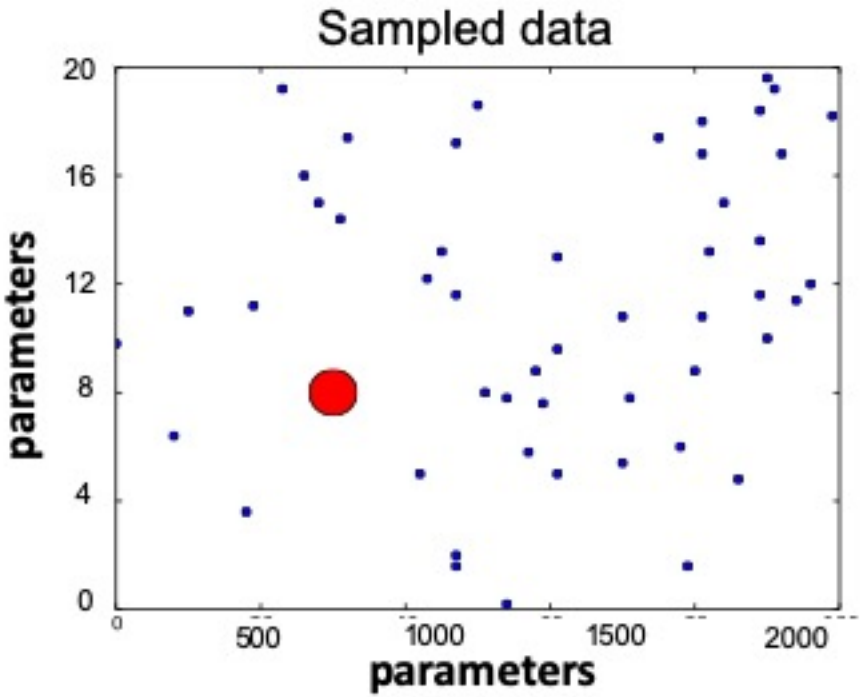
K Nearest Neighbors Model



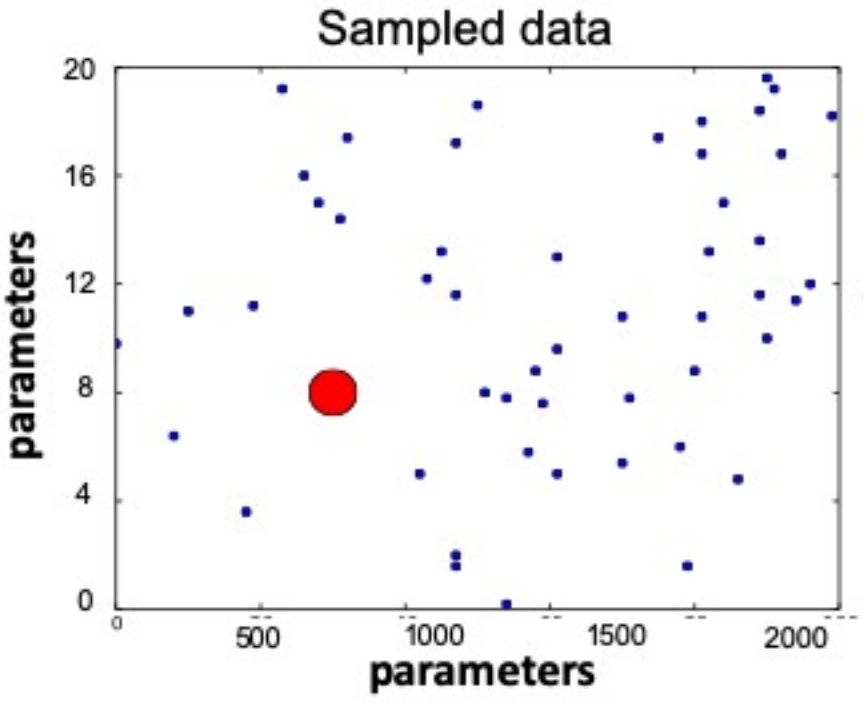
- ***K Nearest Neighbors:*** KNN assigns each point in the testing set a soil moisture that is the weighted average of the soil moisture values of its neighbors

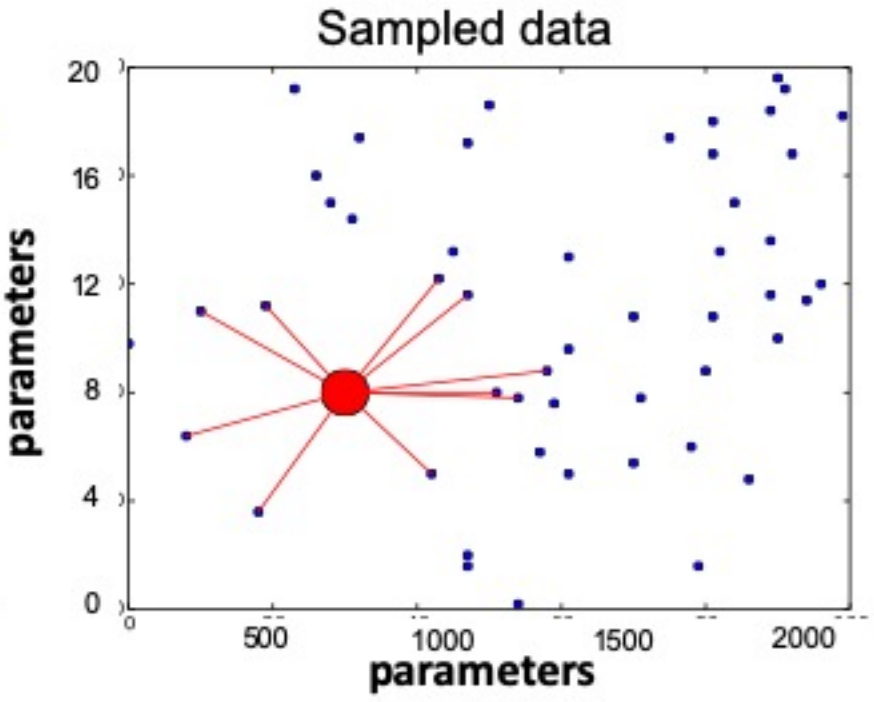
k Nearest Neighbors Model





k Nearest Neighbors Model

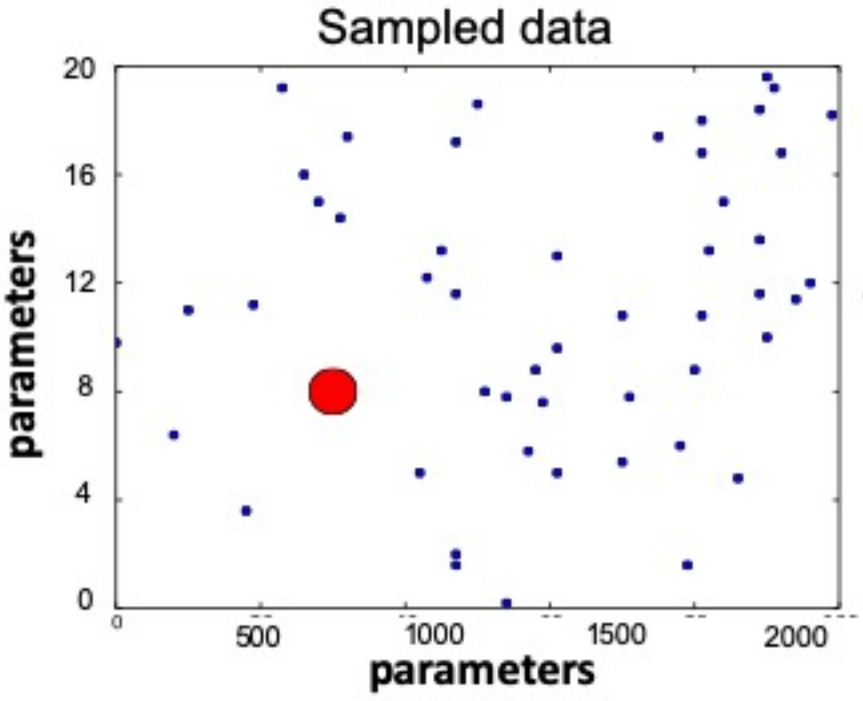


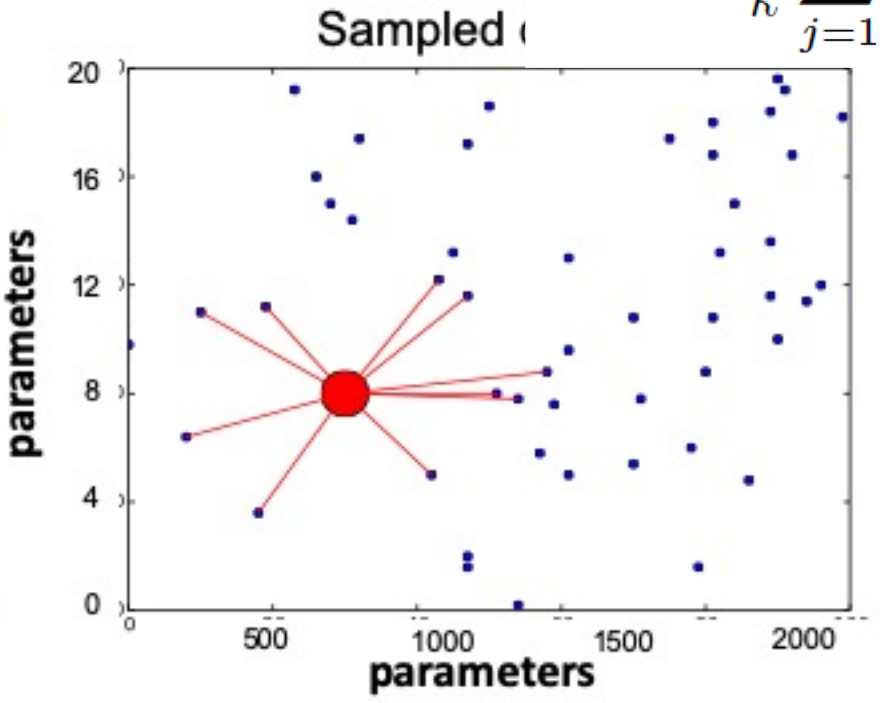


k Nearest Neighbors Model

Polynomial with degree zero

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





k Nearest Neighbors Algorithm

- Import the dataset and curate it – e.g., take care of missing data
- Divide the dataset into disjoined training and test datasets
  - Build model on training data
  - Test model on un-seen data, i.e., test data
- No single way to slit data
  - E.g., splits dataset into 80% train data and 20% test data

k Nearest Neighbors Algorithm

- Split our dataset into two parts, a training and an independent test set

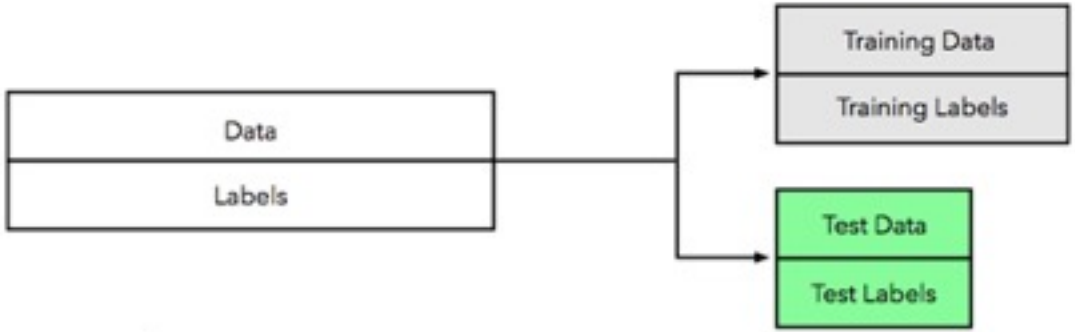


Image from Karl Rosaen
Log http://karlrosaen.com/ml/learning-log/2016-06-20/

k Nearest Neighbors Algorithm

- Apply kNN on training dataset

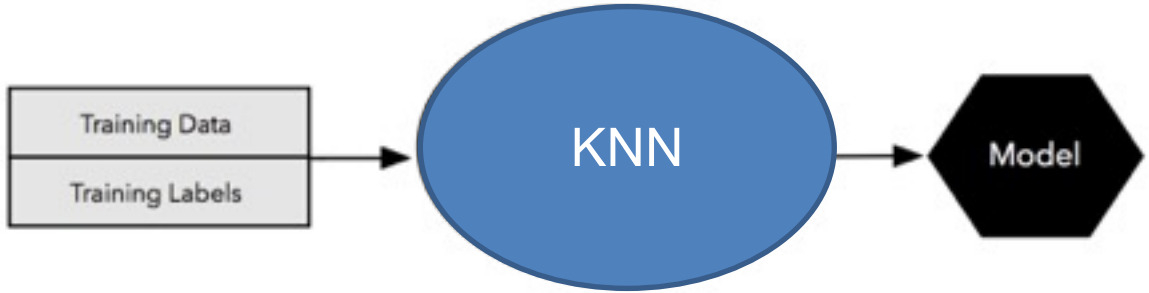


Image from Karl Rosaen
Log http://karlrosaen.com/ml/learning-log/2016-06-20/

k Nearest Neighbors Algorithm

- Test model on un-seen data, i.e., test data

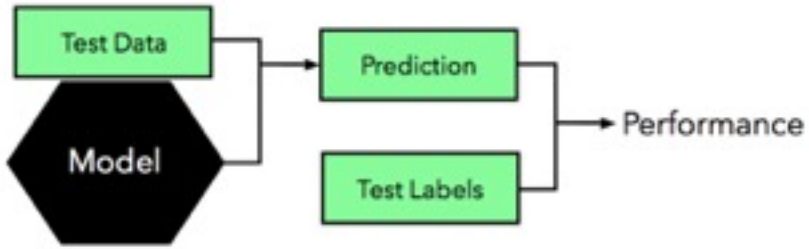
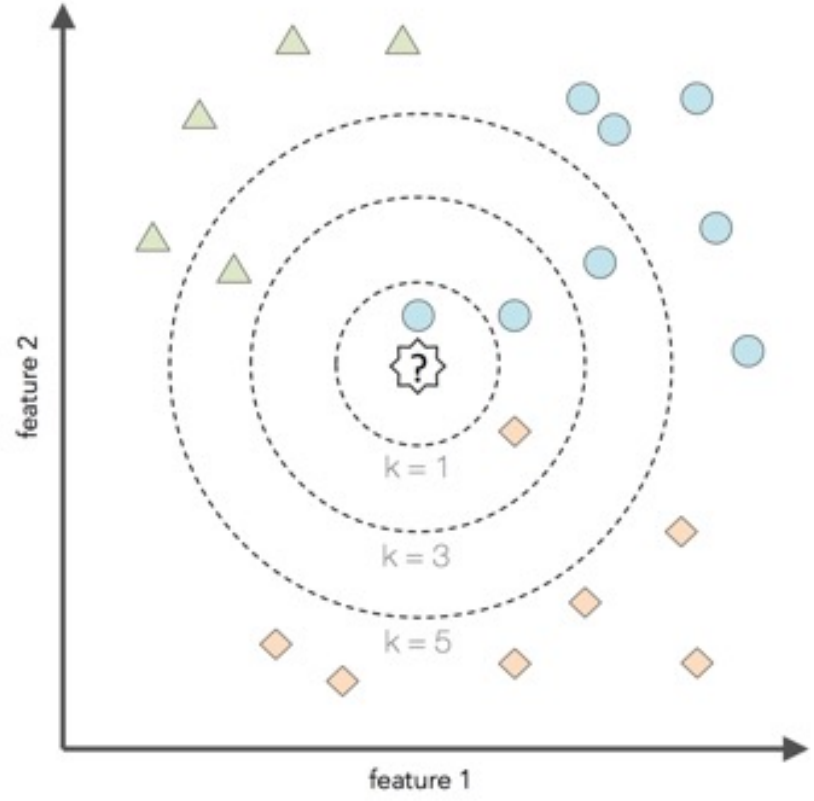


Image from Karl Rosaen
Log http://karlrosaen.com/ml/learning-log/2016-06-20/

Selecting k

- No ideal value for k – select it after testing and evaluation
- Too small: variable, unstable prediction
- Too big: everything like the most probable observation

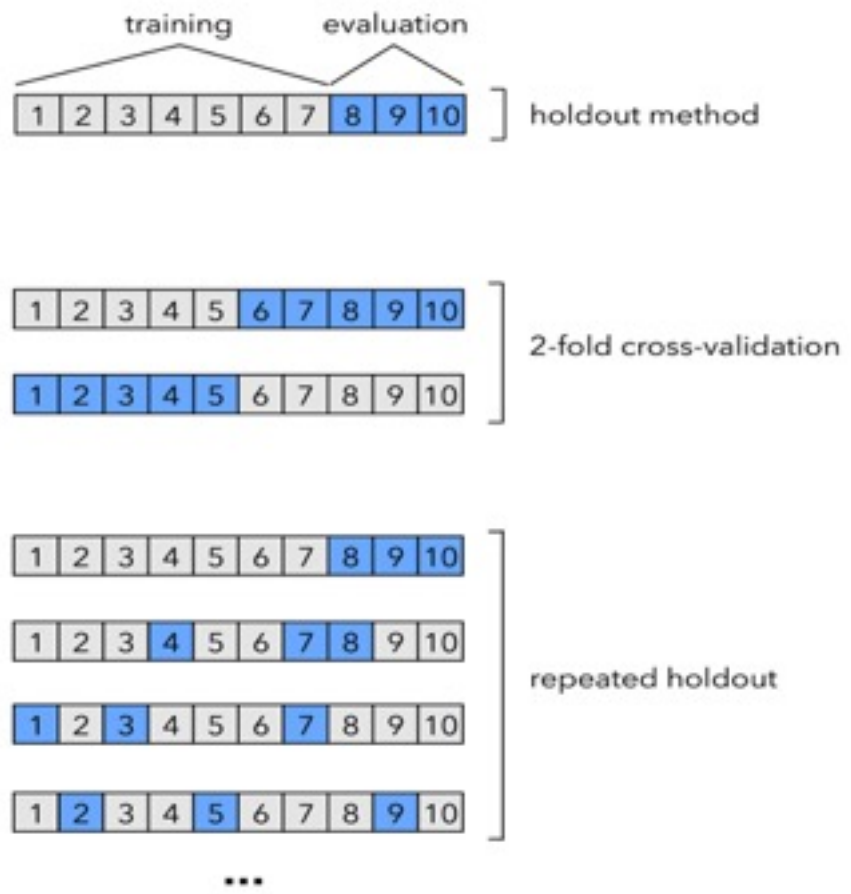


Selecting k

- No ideal value for k – select it after testing and evaluation

Most popular:

- Apply the N-fold cross-validation on training set
- Generate multiple models
- Performance estimates using the test set



N-fold cross validation

No ideal value for k – select it after testing

- Apply N-fold cross validation to select k
  - Randomly partition data into N sets of equal size
  - Build models by using N - 1 of the sets (training sets)
  - Score model by using testing set → possible scores: mean squared error (MSE) of the predictions
- Repeat process N times so that each set serves as testing set

Repeat process with a new random partition of the data into N sets …..

10-fold cross validation

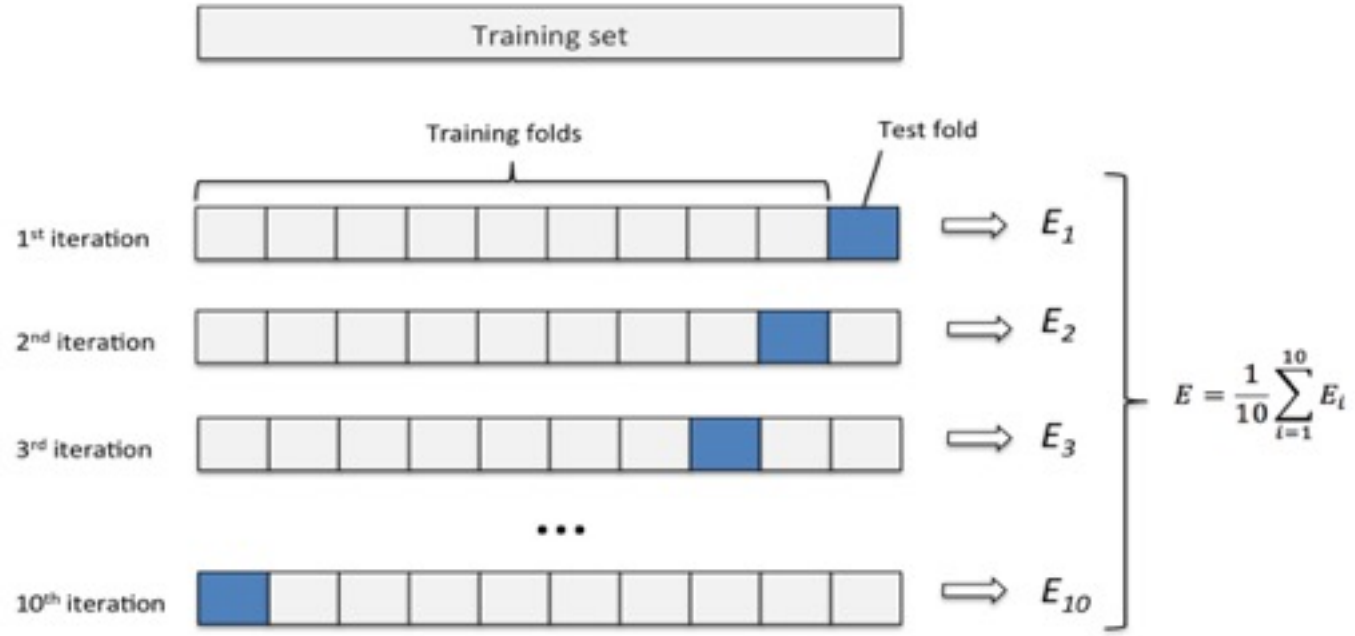
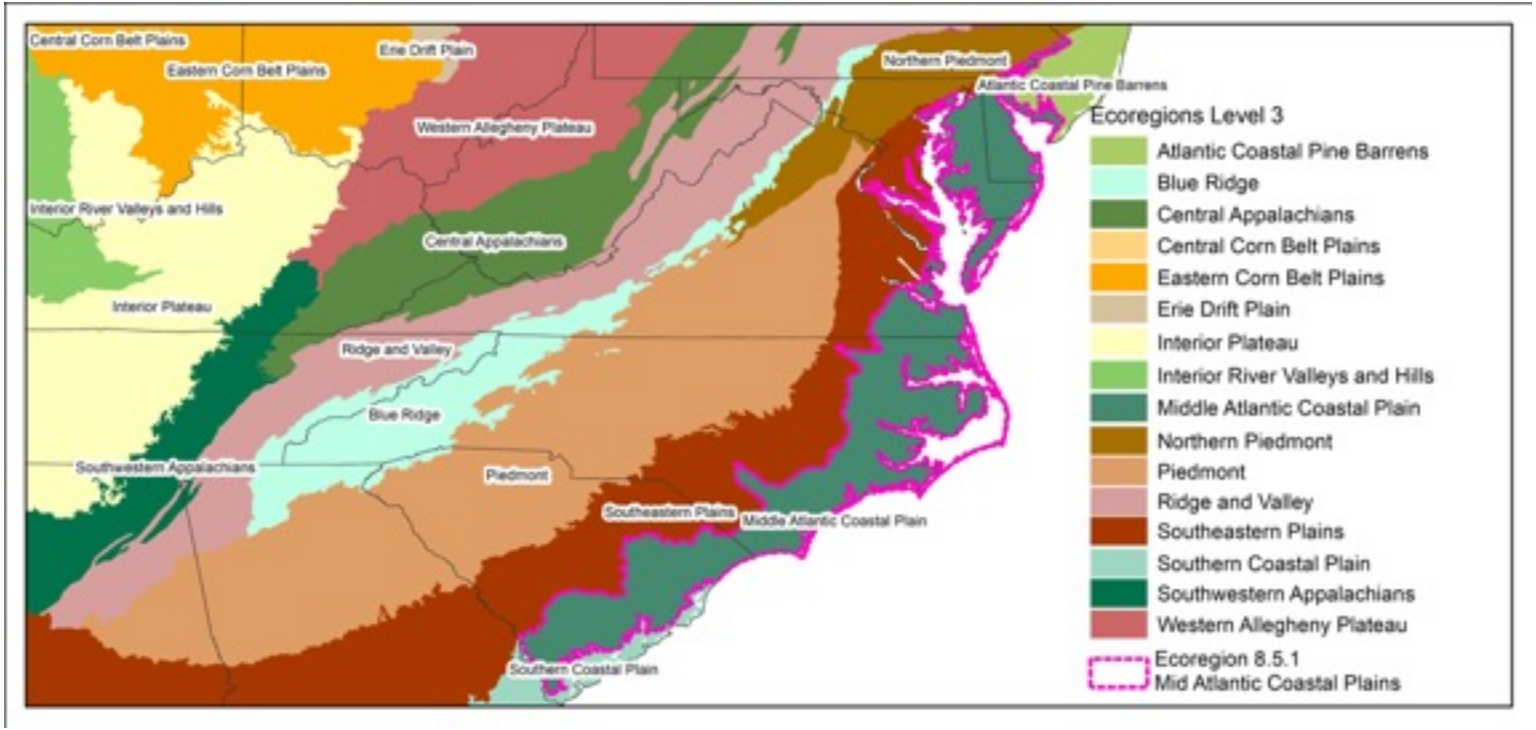


Image from Karl Rosaen
Log http://karlrosaen.com/ml/learning-log/2016-06-20/

Middle Atlantic Coastal Plains

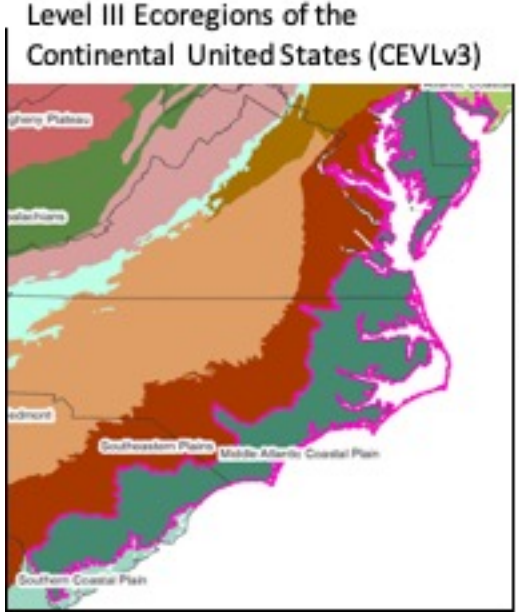


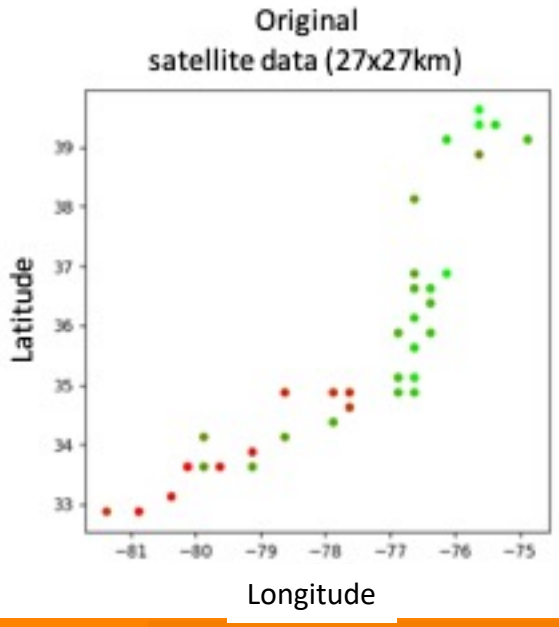
Level III ecoregion 8.5.1 Region with a broad range of moisture ratios

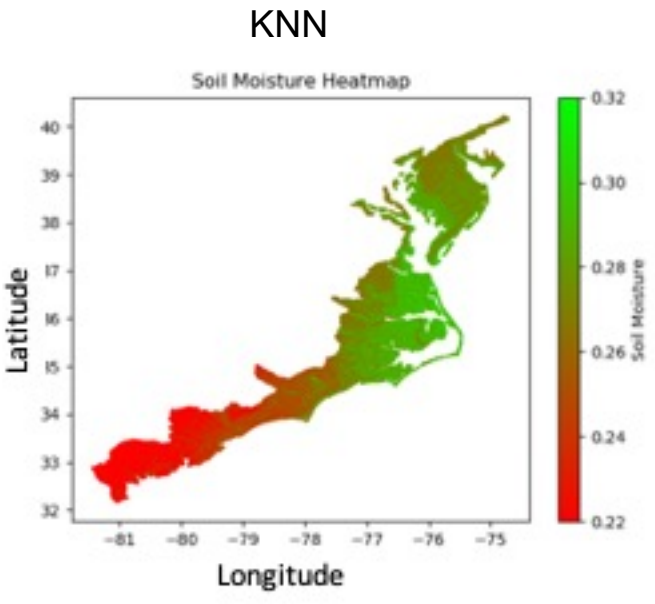
Use case I: from 27x27km to 1x1km

Fine-grained modeling of Mid-Atlantic region in April 2017:

- Terrain parameters: Elevation, Slope, and Wetness Index

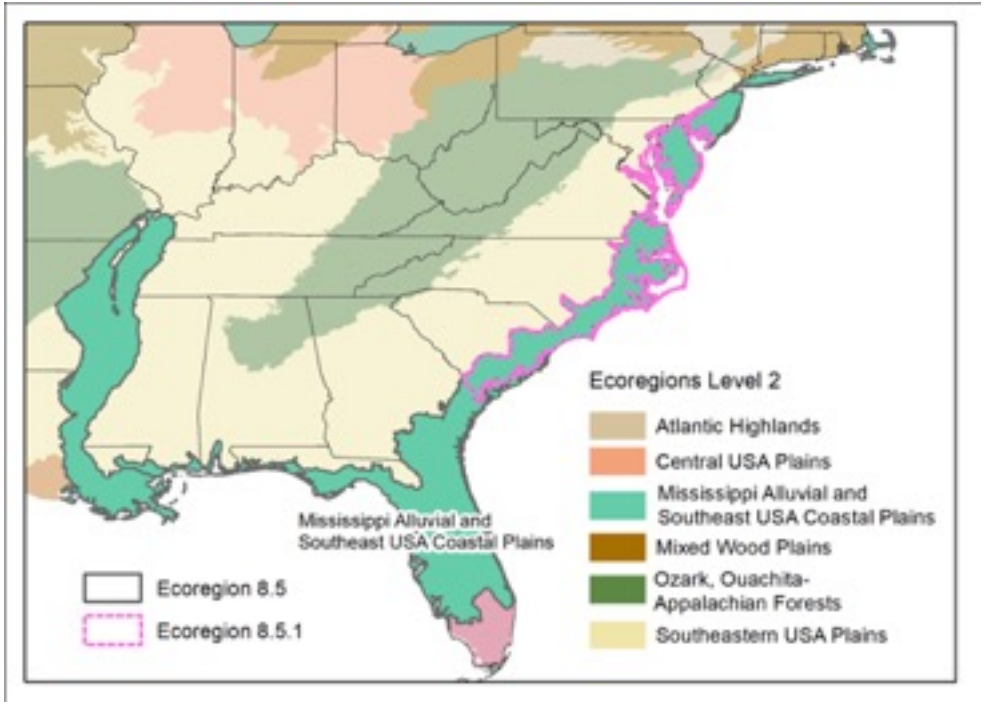




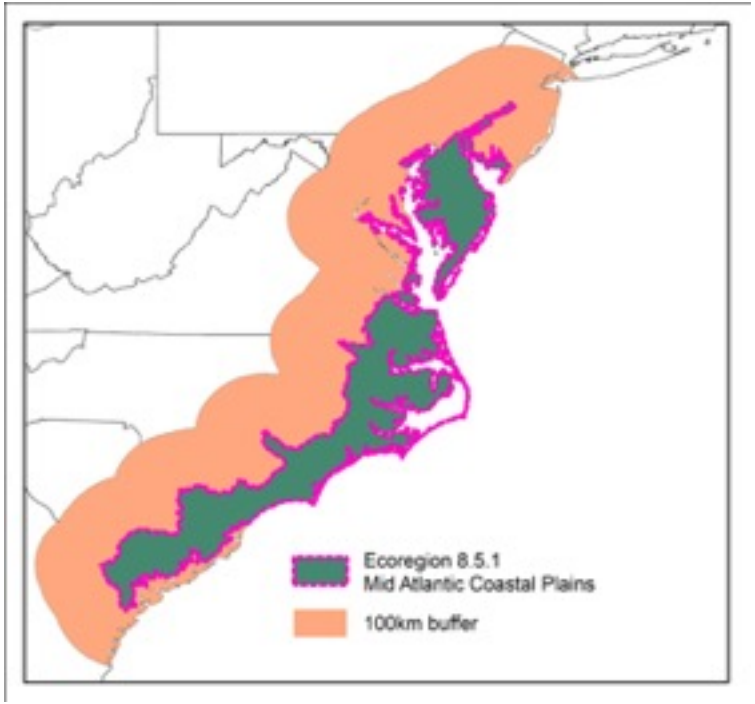


Middle Atlantic Coastal Plains

Level II ecoregion 8.5



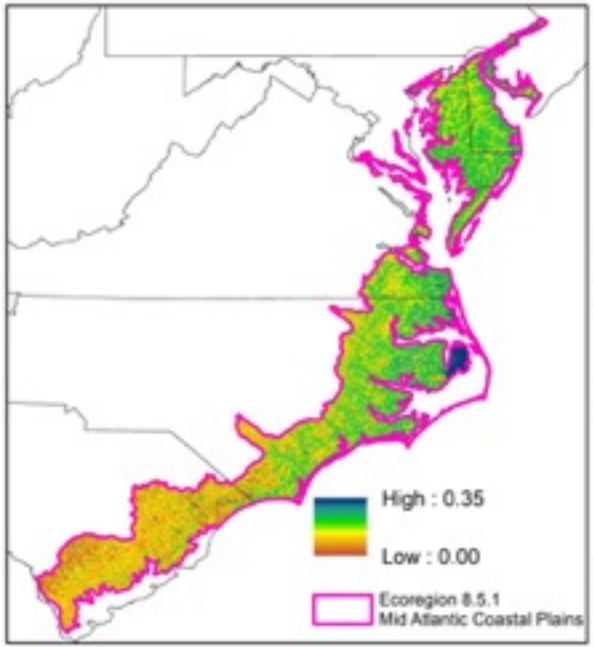
Level III ecoregion 8.5.1 + 100km buffer



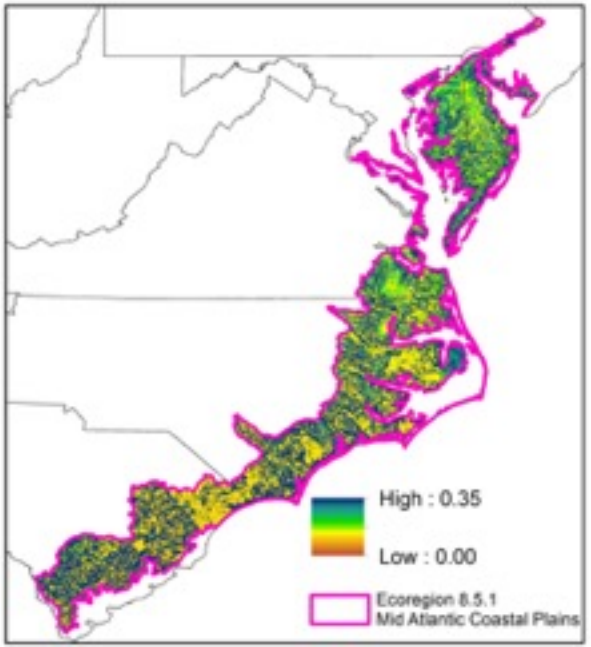
Regions with a broad range of moisture ratios (not in the assignment)

Predictions with different datasets

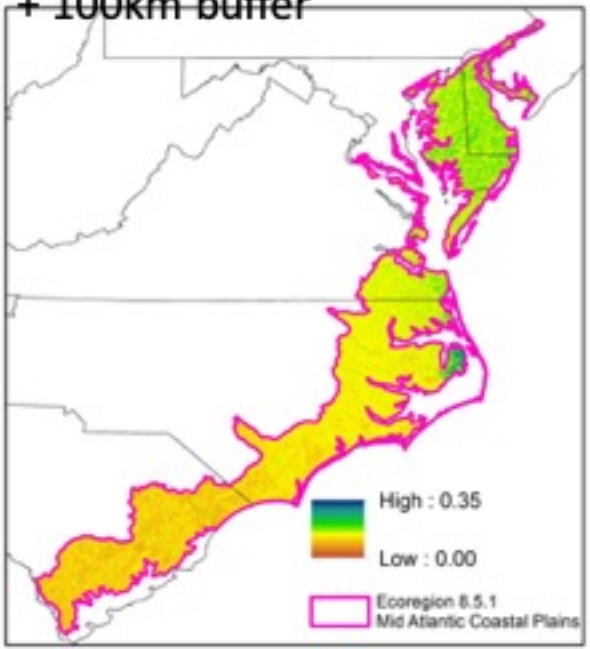
Level III ecoregion 8.5.1



Level II ecoregion 8.5



Level III ecoregion 8.5.1 + 100km buffer



Assignment 9

Introducing pandas

- Walk through some early stages of data downscaling
- Introduce you to a powerful Python package for data analysis: Pandas
- Use Pandas to executed these stages
  - data processing,
  - modeling to generate fine-scale predictions, and
  - visualization

Introducing pandas

- In previous assignments, we used
  - csv library to read data files
  - numpy and pyspark libraries to deal with missing values
- Both functionalities are readily available with pandas

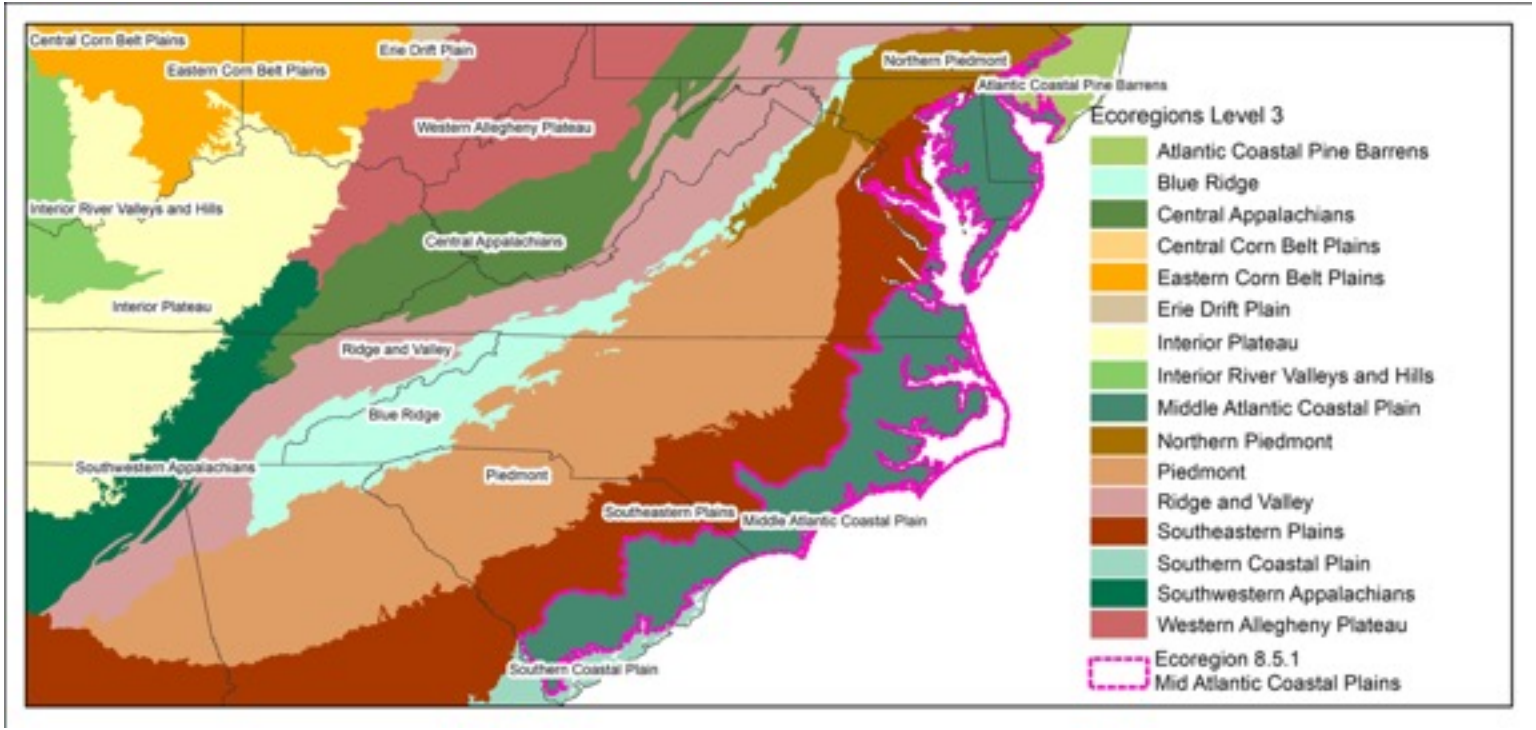
Assignment 9

- Preprocessing data (different from dietary data):
- Problem 1: Remove all monthly columns from a given data frame except for a specified month
- Problem 2: Drop the rows that have an NA in that column
- Problem 3: Use KNN to predict data on the region of interest
- Problem 4: Create heatmaps for the training and prediction soil moisture values
  - Course-grained map: Original Soil Moisture Heatmap
  - Fine-grained map: Predicted Soil Moisture Heatmap

Data features and preparation

- Process data in Pandas DataFrames to prepare it for modeling
- Data for the Mid Atlantic Coastal Plains, a North American ecoregion containing the state of Delaware
- Subset of data for a single year (2016)
- Two comma-separated text files:
  - "Delaware_train.csv" with 29 columns
  - "Delaware_eval.csv" with 17 columns
- Initial two columns: x (longitude) and y (latitude) coordinates
- Final fifteen columns: 15 topographic parameters.
- Twelve columns (of the "train" file ONLY): monthly soil moisture averages for 2016 for the approximately 27 km x 27 km pixel with centroid at the given coordinates

Middle Atlantic Coastal Plains



Level III ecoregion 8.5.1 Region with a broad range of moisture ratios

Data predictions

- Feed the processed data into a pre-prepared modeling script to produce downscaled soil moisture predictions
- Coarse-resolution "train" data is used to generate a model
- Model is evaluated on the fine-resolution "eval" data

Data visualization

- Take advantage of Pandas' integration with matplotlib to create heatmaps for visually comparing your downscaled soil moisture product to the original data

Project

Project (I)

Motivation Describe the motivation of your work. To build the motivation, you can answer these questions:

- What is the problem you are tackling?
- How is the problem solved today?
- Write a paragraph of 200 - 300 words

Contributions List between 2 and 4 contributions of your work. Contributions are bullet points that define your solution. E.g.,

- We build a system that ....
- We validate the system accuracy by ....
- We measure the performance of the system by ...
- Write a section of 150 - 200 words

Project (II)

Tests List the type of tests (measurements) you will perform. E.g.,

- What are your metrics of success?
- Where do you run your tests?
- What tests do you perform?
- How many times do you run each test?
- What do you measure?
- Write a section of 250 - 350 words.

Reading

Paper

- D. Rorabaugh, M. Guevara, R. Llamas, J. Kitson, R. Vargas, and M. Taufer. SOMOSPIE: A Modular SOil MOisture SPatial Inference Engine based on Data Driven Decisions. IEEE eScinece Conference, 2019