PolluteMeNot[AI].life

Joseph Gaustad
Madhukar Reddy
Shobha Sankar
Radhika Satapathy

MIDS FINAL CAPSTONE PROJECT PRESENTATION

School of Information, UC Berkeley

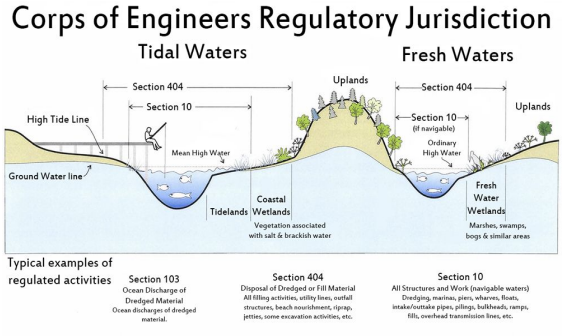
April 15, 2021

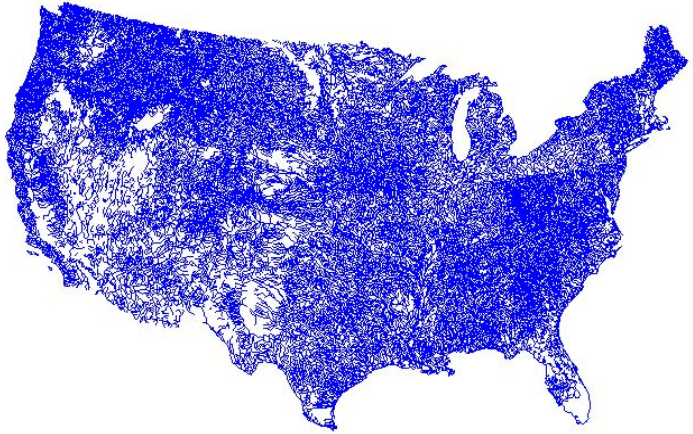
Clean Water Act (CWA) was passed in 1972 and governs pollution of US surface waters…

Permits for Clean Water
NPDES

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

US Army Corps of Engineers®





"Waters of the United States"
WOTUS


Farmer


Fossil Fuel Producer


Condominium Builder

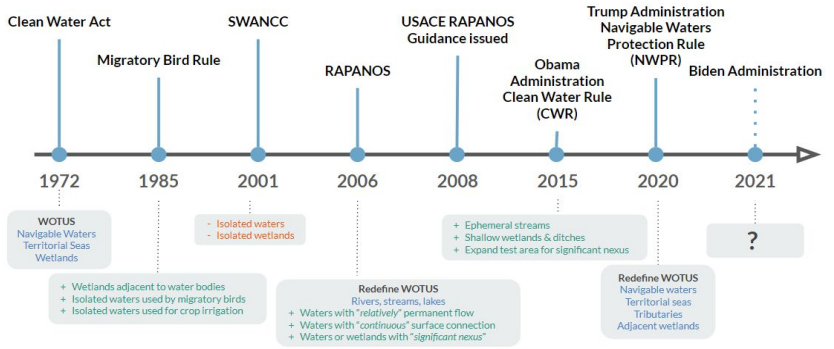

Real Estate Developer


Home Owner

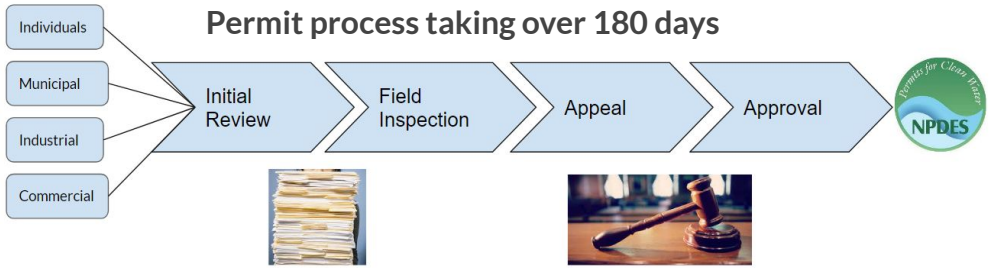

Water Treatment Plant

CWA has undergone numerous changes and regulation has become a challenge…

Legislative and Judicial changes over the years



Permit process taking over 180 days



- Changing Jurisdictions
- Unclear Definitions
- No Maps
- Outdated Data
- Manual Method

Can technology help determine which waters are under jurisdiction?

Joseph S. Shapiro
Associate Professor
UC Berkeley
Dept. of Agriculture & Resource Economics

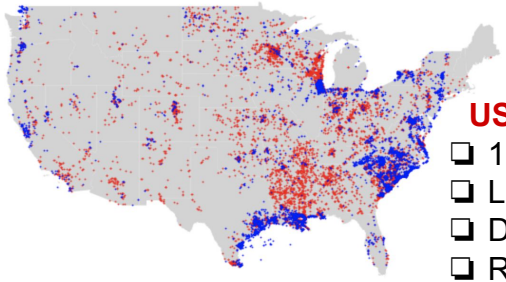
Manuela Girotto
Assistant Professor
UC Berkeley
Dept. of Environmental Science, Policy & Management

David A. Keiser
Associate Professor
UMass Amherst
Dept. of Resource Economics

RESEARCH QUESTION

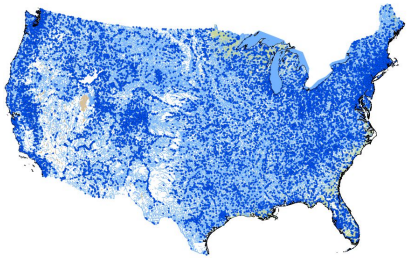
"**Can we** develop an AI based technique which can consider satellite imagery and geospatial data, and predict whether a given site has jurisdictional waters per the Clean Water Act?"

US Army Corps JDs & Geospatial Data



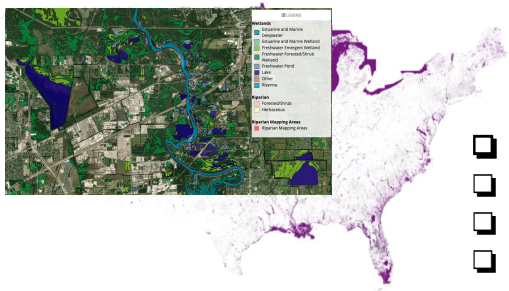
US Army Corps Database

- ❑ 15K CWA Determinations
- ❑ Location
- ❑ Decision Criteria
- ❑ RAPANOS



NHD

- ❑ **USGS & US EPA**
- ❑ Geospatial
- ❑ Surface Water Network
- ❑ Rivers, Lakes, Streams, Canals…
- ❑ Most up-to-date



NWI

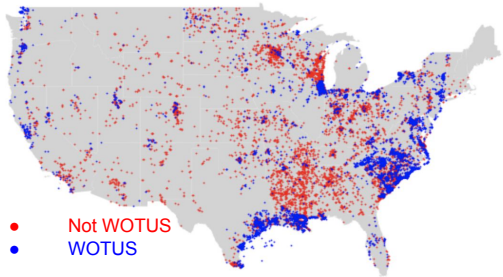
- ❑ **USFWS (Fish & Wildlife)**
- ❑ Geospatial
- ❑ National Wetlands Inventory (NWI)
- ❑ Abundance, Characteristics

SSURGO

- ❑ **USDA (Agriculture)**
- ❑ Geospatial
- ❑ Soil Characteristics
- ❑ Water content

Data Sources - US Geospatial Sources

Army Corps Decisions

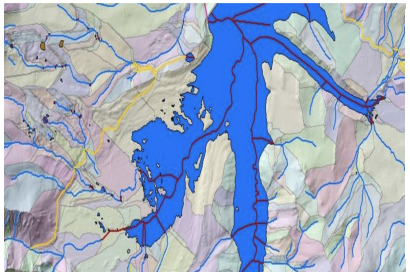


- 15K Decisions
- Latitude / Longitude
- Decision Criteria
- RAPANOS/ CWR / NWPR
- ...

Soil Survey - USDA

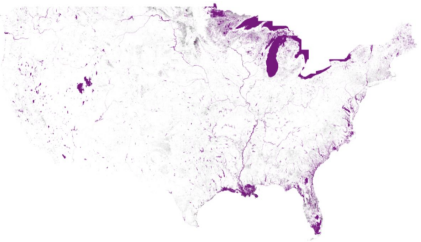
- Hydrology group
- Hydric Classification
- Flood Frequency
- Water table depth
- ...

Hydrography - USGS



- Flowline shapefile
- Waterbody shapefile
- District Level
- Medium Resolution
- ...

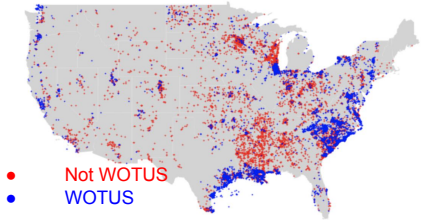
Wetlands Inventory-NWI



- Waterbody Shapefiles
- Rivers/Streams Shapefiles
- Classification
- Incomplete
- ...

Data Engineering

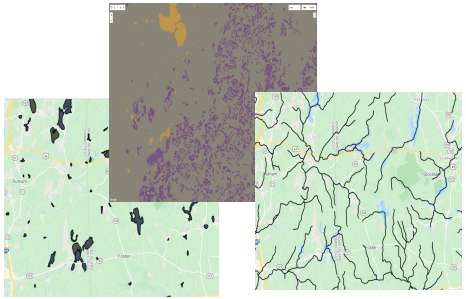
Army Corps Decisions



Sentinel-2 Remote Sensing

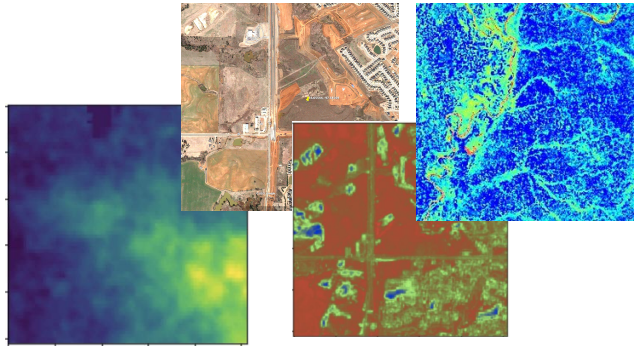
Geospatial Dataset

- Location Specific
  - Soil Features
  - Hydrography Features
  - Wetland Features



Satellite Imagery

- Location Specific Images
  - Sentinel-2 (10 m resolution)
    - RGB
    - Water & Vegetation indices
  - SRTM (30 m resolution)
    - Slope & Elevation
  - JRC (30 m resolution)
    - Seasonality & Transition



Satellite Imagery

Sentinel-2 bands	Spatial resolution (m)
Band 1 – Coastal aerosol	60
Band 2 – Blue	10
Band 3 – Green	10
Band 4 – Red	10
Band 8 – NIR	10
Band 11 – SWIR	20

- 256 x 256 pixel patch
- Hi-res: 10 m / pixel
  - 2.5 KM x 2.5 KM
- Lo-res: 100 m / pixel
  - 25 KM x 25 KM

Sentinel-2 (10 m)

RGB

Near InfraRed

Short-wave IR

Water & Vegetation indices

mndvi

mndwi

gmndwi

SRTM (30 m)

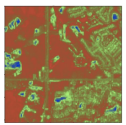
Elevation

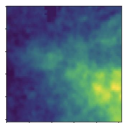
Slope

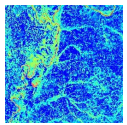
JRC (30 m)

Transition

Seasonality







Data Engineering

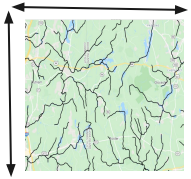
Sentinel-2: 10m resolution
- RGB
- Water & Vegetation

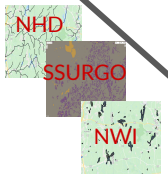
SRTM & JRC : 30m resolution
- Seasonality & Transition
- Elevation & Slope

Google Earth Engine

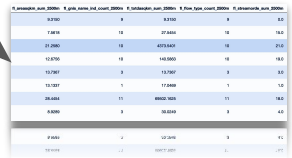
- 15K Locations
- Soil Features (SSURGO)
- Hydrography Features(NHD)
- Wetlands Features (NWI)

200 x 200 m patch
1000 x 1000 m patch
2500 x 2500 m patch

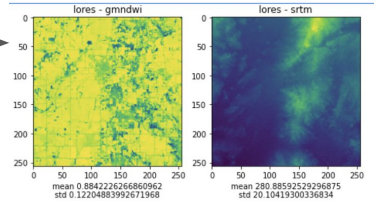


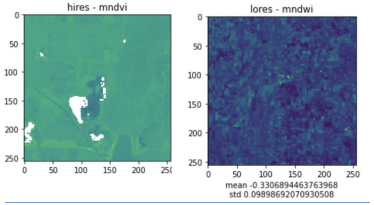


Cloud Storage



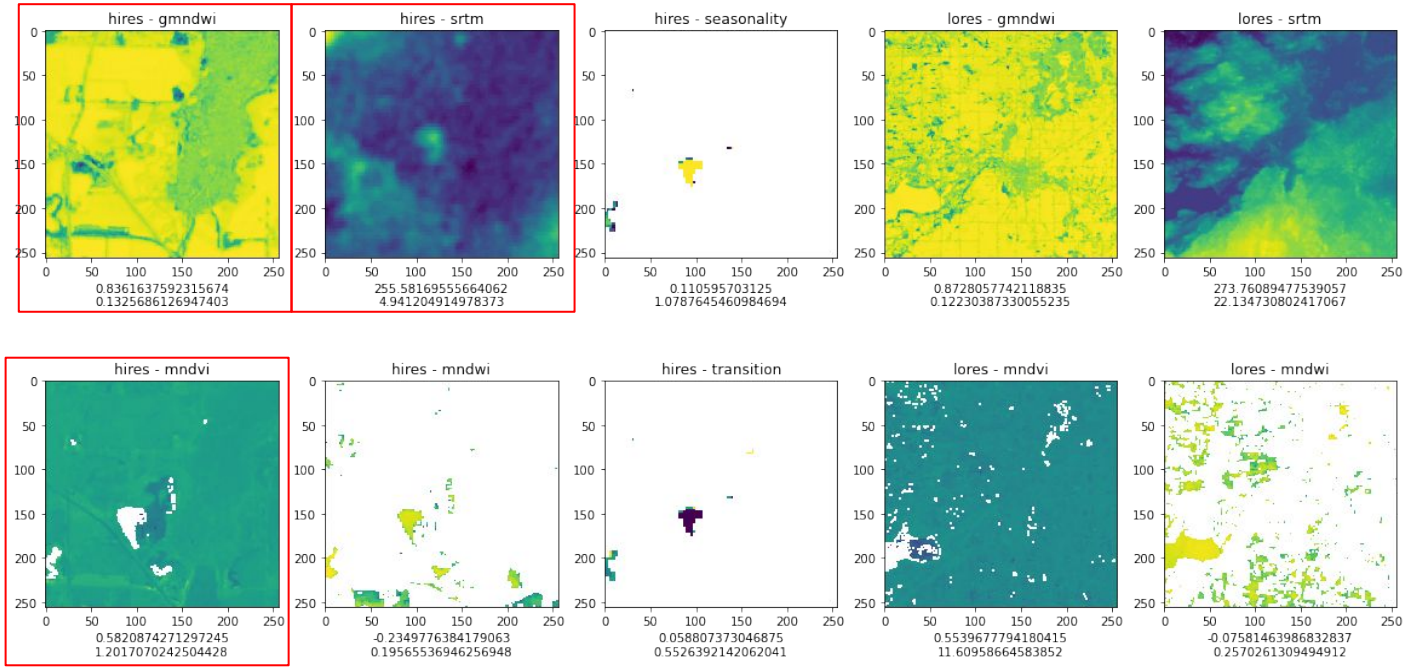
- MNDVI, MNDWI, GMNDWI, SRTM, Seasonality, Transition
- 6 Hi-res + 4 Lo-res images
- 10 images / location
- 150K images
- 4 weeks @ 20min/image



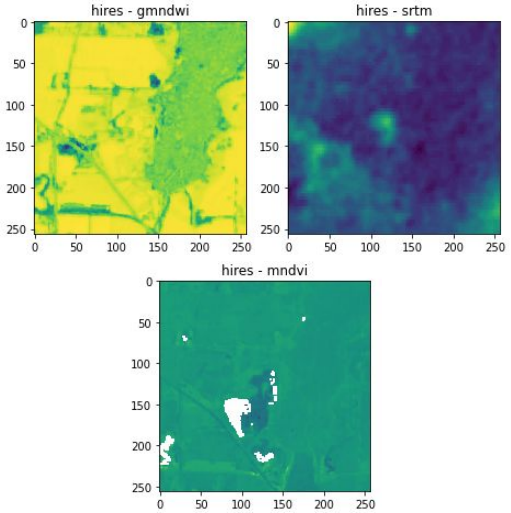


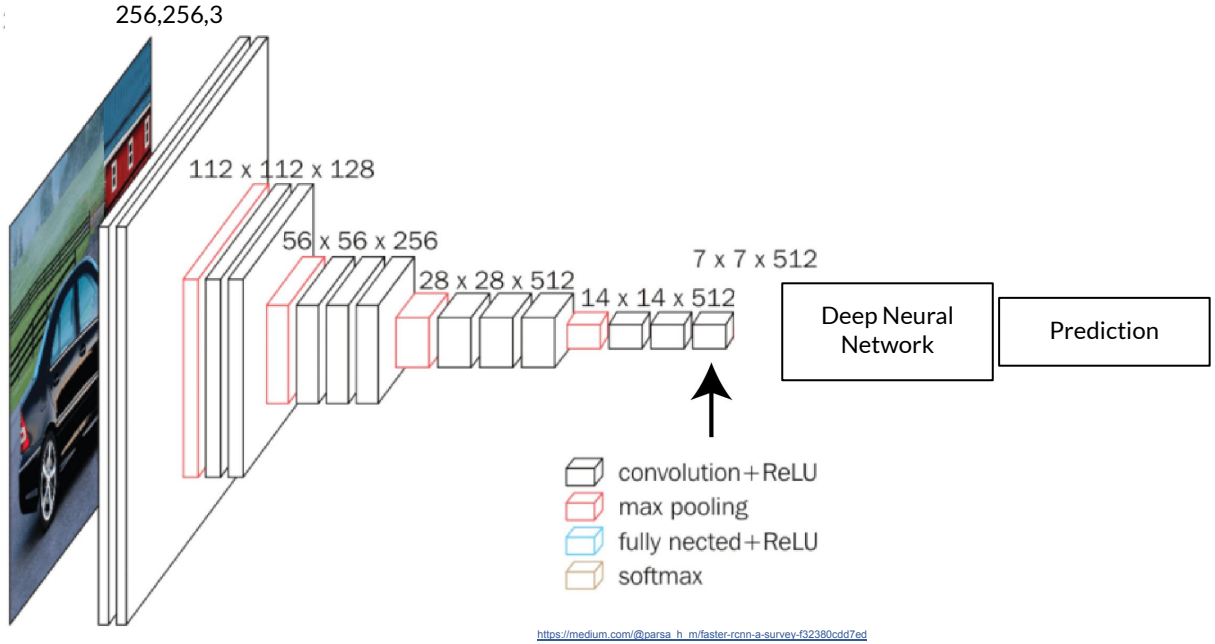
Layers of Data

hires: 2.5km X 2.5km
lores: 25km X 25km



Model: Transfer Learning and CNNs





https://medium.com/@parsa_h_m/faster-rcnn-a-survey-f32380cdd7ed

CNN Model: Results

ROC_AUC: 0.657

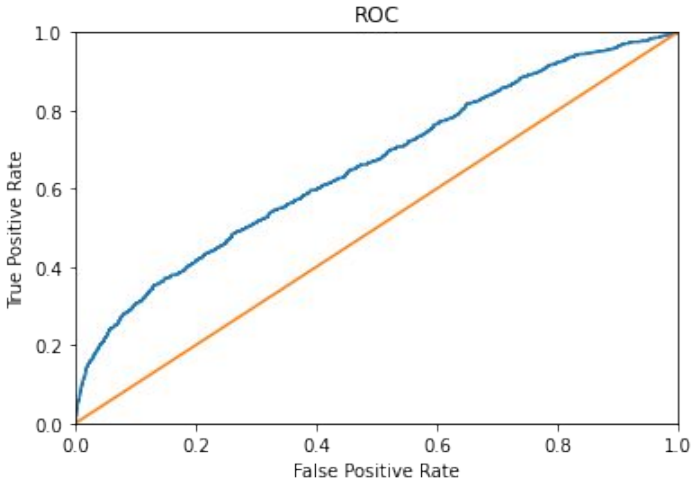
Avg. Precision: 0.69

Balanced Accuracy: 0.591

Areas for more investigation:

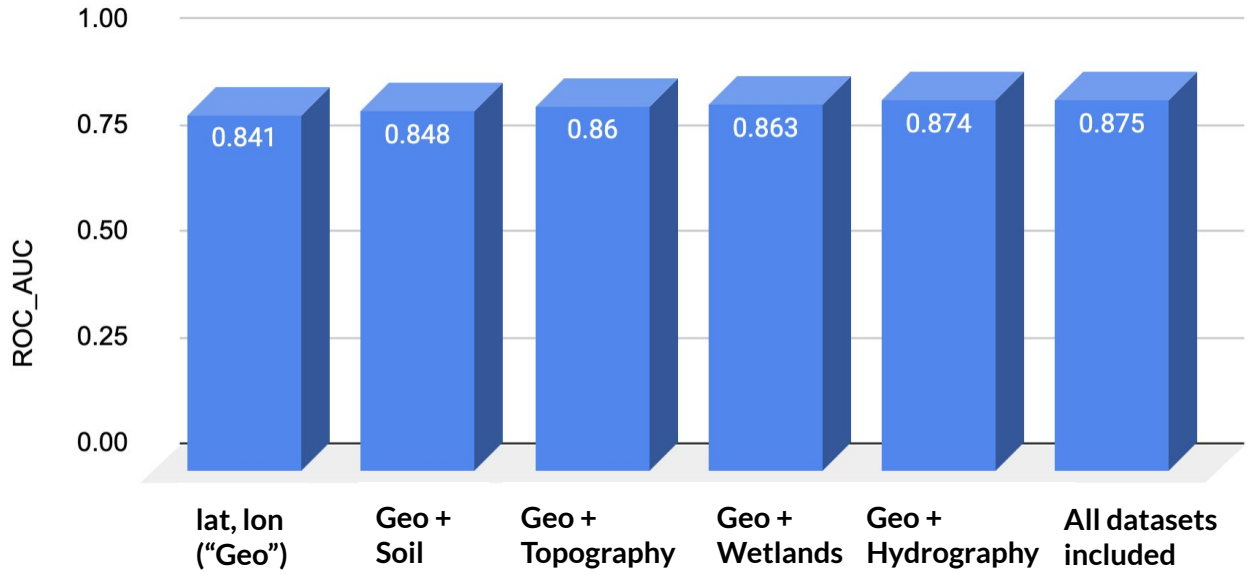
- Increase resolution
- Increase count of data points
  - Train region specific models

```
Confusion Matrix
[[2605  157]
 [1163  367]]
```



	precision	recall	f1-score	support
0	0.69	0.94	0.80	2762
1	0.70	0.24	0.36	1530
accuracy			0.69	4292
macro avg	0.70	0.59	0.58	4292
weighted avg	0.69	0.69	0.64	4292

Light GBM Performance on Tabular Datasets

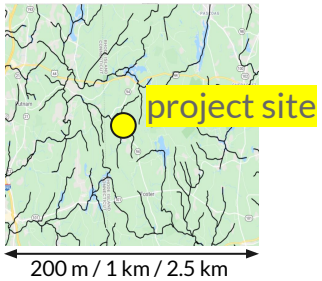


*200mX200m, 1kmX1km and 2.5kmX2.5km satellite image patch sizes

Top Predictors

Geographical Information (lat, lon, district)	ROC	Top Predictors*
Included	0.87	lat, lon (elevation, slope)[a] (seasonality, recurrence)[b] (soil - water depth)[c] (extent of waterbodies)[b]
Not included	0.83	(elevation, slope)[a] (seasonality, recurrence)[b] (soil - water depth)[c] (extent of waterbodies)[b]

[a]200m or 1000m patch size
[b]2.5km patch size
[c]point information



Model Performance by Jurisdictional District

Great performance in most districts

Weak performance is some key districts

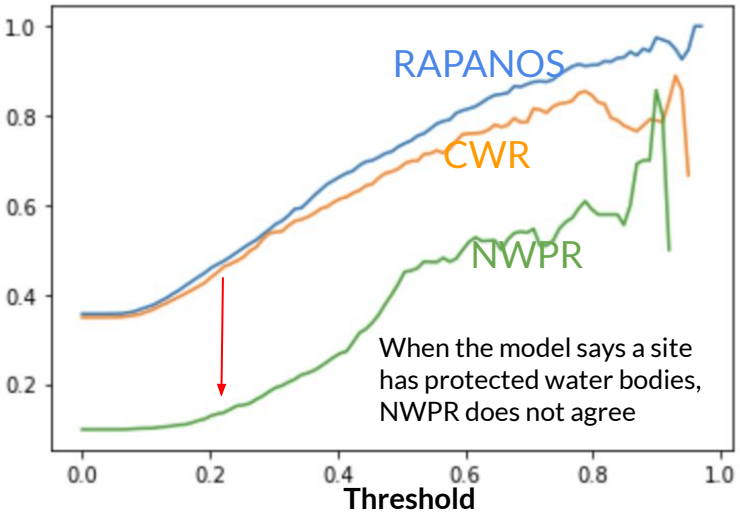
District	Count	WOTUS	ROC
Walla Walla	13	0.23	1.000
Honolulu	27	0.30	1.000
St. Louis	4	0.75	1.000
Memphis	120	0.08	0.991
Mobile	346	0.18	0.979
St. Paul	1160	0.07	0.953
Rock Island	64	0.69	0.944
Galveston	834	0.63	0.914
Seattle	88	0.70	0.878
Los Angeles	159	0.14	0.826
Alaska	359	0.22	0.815
New Orleans	819	0.48	0.810
Norfolk	510	0.47	0.789
Wilmington	1074	0.72	0.774
Little Rock	139	0.28	0.765
Savannah	105	0.18	0.758
Buffalo	220	0.59	0.734
San Francisco	69	0.78	0.733
New York	103	0.72	0.730

District	Count	WOTUS	ROC
Chicago	462	0.58	0.730
Kansas City	217	0.13	0.708
Omaha	408	0.21	0.698
Charleston	1687	0.37	0.691
Detroit	142	0.44	0.690
Baltimore	115	0.66	0.686
Tulsa	109	0.26	0.684
Vicksburg	762	0.01	0.657
Portland	76	0.50	0.636
Huntington	256	0.08	0.633
Philadelphia	76	0.39	0.626
Louisville	371	0.13	0.586
Nashville	104	0.09	0.513
Sacramento	232	0.32	0.496
Jacksonville	66	0.59	0.467
Albuquerque	97	0.06	0.400
Pittsburgh	26	0.50	0.250
New England	22	0.36	0.100
Fort Worth	16	0.62	NaN

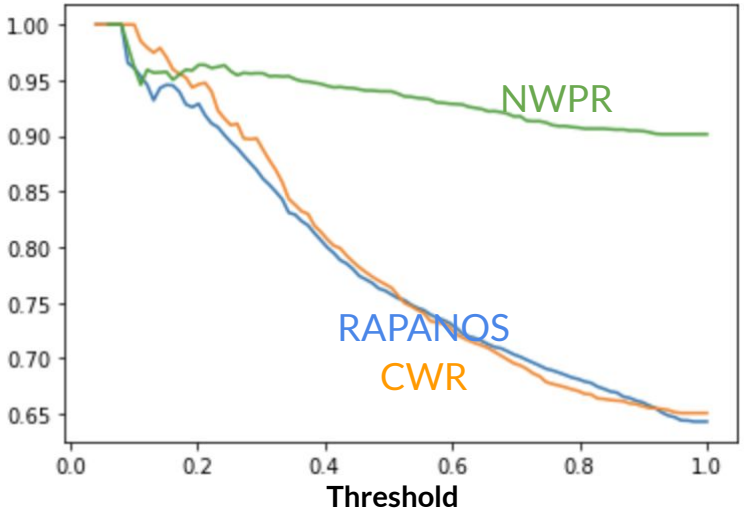
WOTUS Protection: RAPANOS > CWR > NWPR

2006 2015 2020

Positive Predictive Value



Negative Predictive Value



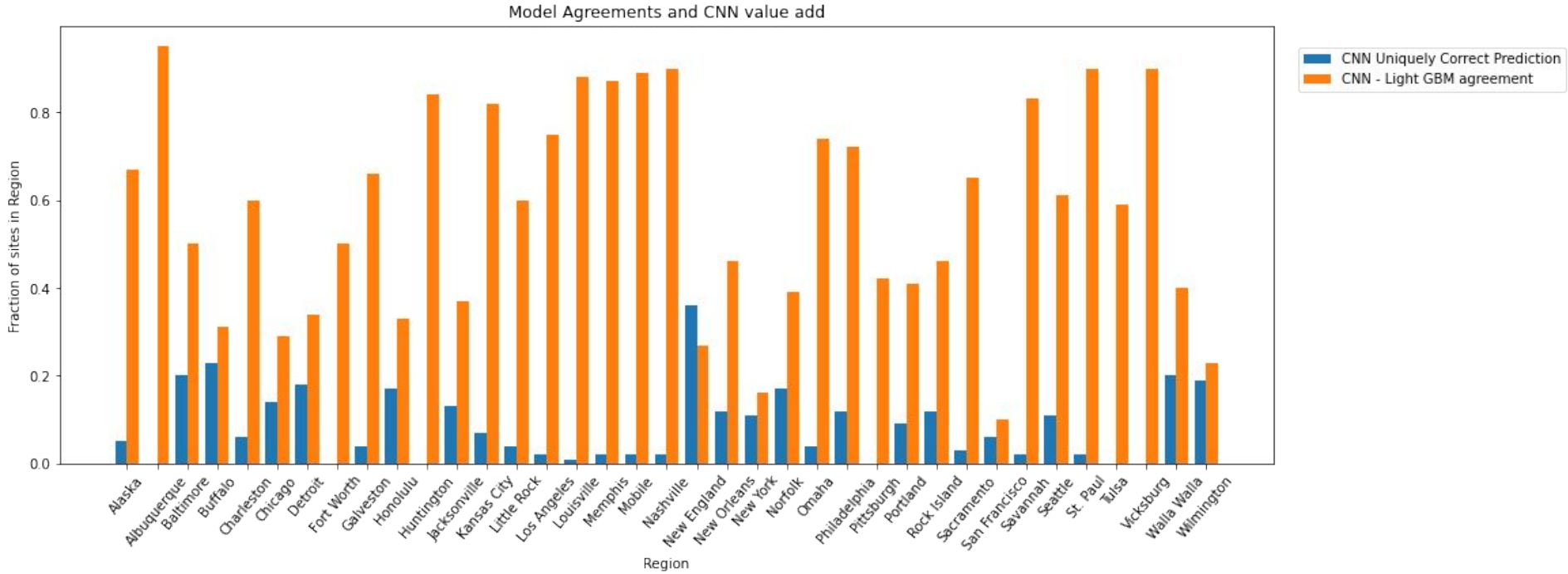
Value Add by CNN Model

- 7.5% of the site jurisdictions were uniquely predicted correct by the CNN model.

district	CNN unique correct	GBM unique correct	Model Agreement	Count
Alaska	0.05	0.12	0.67	130
Albuquerque	0.00	0.02	0.95	44
Baltimore	0.20	0.13	0.50	30
Buffalo	0.23	0.30	0.31	84
Charleston	0.06	0.08	0.60	585
Chicago	0.14	0.36	0.29	167
Detroit	0.18	0.16	0.34	44
Fort Worth	0.00	0.00	0.50	2
Galveston	0.04	0.16	0.66	293
Honolulu	0.17	0.17	0.33	6
Huntington	0.00	0.06	0.84	94
Jacksonville	0.13	0.17	0.37	30
Kansas City	0.07	0.02	0.82	82
Little Rock	0.04	0.09	0.60	47
Los Angeles	0.02	0.03	0.75	64
Louisville	0.01	0.02	0.88	116
Memphis	0.02	0.05	0.87	61
Mobile	0.02	0.05	0.89	137
Nashville	0.02	0.00	0.90	48

district	CNN unique correct	GBM unique correct	Model Agreement	Count
New England	0.36	0.18	0.27	11
New Orleans	0.12	0.19	0.46	303
New York	0.11	0.54	0.16	37
Norfolk	0.17	0.29	0.39	201
Omaha	0.04	0.06	0.74	154
Philadelphia	0.12	0.12	0.72	25
Pittsburgh	0.00	0.00	0.42	12
Portland	0.09	0.21	0.41	34
Rock Island	0.12	0.35	0.46	26
Sacramento	0.03	0.01	0.65	80
San Francisco	0.06	0.39	0.10	31
Savannah	0.02	0.02	0.83	46
Seattle	0.11	0.14	0.61	28
St. Paul	0.02	0.05	0.90	457
Tulsa	0.00	0.03	0.59	34
Vicksburg	0.00	0.09	0.90	318
Walla Walla	0.20	0.00	0.40	5
Wilmington	0.19	0.49	0.23	429

Comparative Analysis: CNN Vs Light GBM



7.5% positive predictions were uniquely made by CNN

Spectral imagery is helpful

Conclusion

*"**Can we** develop an AI based technique which can consider satellite imagery and geospatial data, and predict whether a given site has jurisdictional waters per the Clean Water Act?"*

Yes, AI technology shows promise for CWA determinations

Near Term Research

- Higher resolution imagery
- Image pre-processing
- Time-series spectral indices

Longer Term Goals

- Maps
- Consolidated Datasets
- Regulatory Platforms

Let's ensure cleaner water for all!

Questions?

For more information:

http://bit.ly/PolluteMeNot

Individual Contributions - Joe

- Joe
  - Domain knowledge research and met with stakeholders
  - Contributed to extracting and parallelizing the Google Earth Engine data
  - Setup AWS infrastructure and development environment with docker and EFS for all modeling work, including selection of tensorflow as the platform of choice.
  - Image EDA analysis, including developing an ingestion script from GCP (google cloud) to AWS EFS
  - EDA work and investigation to preprocessing image data from rasters.
  - Identification and experimentation of transfer learning opportunities.
  - CNN architecture experimentation and setup including model verification on a known problem like the flower image data set.
  - Investigation and work on augmenting the training data set
  - CNN and Light GBM model comparison and analysis for identifying the value add from the CNN model.

Individual Contributions - Radhika

- Radhika
  - Domain research and problem definition
  - Leading meetings with subject matter experts
  - State of the art literature review for understanding modeling techniques and feature selection
  - Jurisdictional Determinations (JDs) - understanding, EDA and initial visualizations
  - GEE code validation for exporting spectral images, JRC and SRTM; code dev for exporting RGB images
  - External data acquisition (soil data from USDA), understanding NHD
  - Model Building (XGB)
    - Feature selection
    - Pre-processing
    - Choice of sampling technique
    - Class balancing
    - Hyperparameter tuning
    - Ensembling methods (voting, blending, bagging)
    - Results evaluation - metrics and charts
  - Analysis
    - False Negatives
    - Patch Size Effects
    - Region specific models
  - Informational website framework development
  - Slide deck creation & storyline

Individual Contributions - Shobha

- Shobha
  - Domain knowledge acquisition
  - Metadata EDA Visualization & Preprocessing
  - Hydrology data Acquisition & Cleanup (NHD)
  - GEE Image generation
  - Image pre-processing
  - Modeling, Ensembling techniques (Blending, Bagging, Voting)
  - Feature selection, Hyper parameter tuning (Random Forest)
  - Presentation deck

Individual Contributions - Madhukar

- Madhukar
  - Point of contact with professors
  - Successful and timely execution of all aspects of the project owned, and filled gaps that arose
  - Focus on collaboration with external stakeholders professors, instructors, MIDS alumni and NLP team as and when needed
  - Literature survey, research statement refinement, EDA
  - Dataset downloads, gained expertise of GEE API, data ingestion into GEE, features and images extraction from GEE
  - Developed code to extract all the above information
  - Model building, feature engineering, feature selection, hyperparameter tuning
  - Defining the metrics to be used for analysis of obtained results

APPENDIX

Best performing model: Light GBM

ROC_AUC: 0.875 ± 0.009

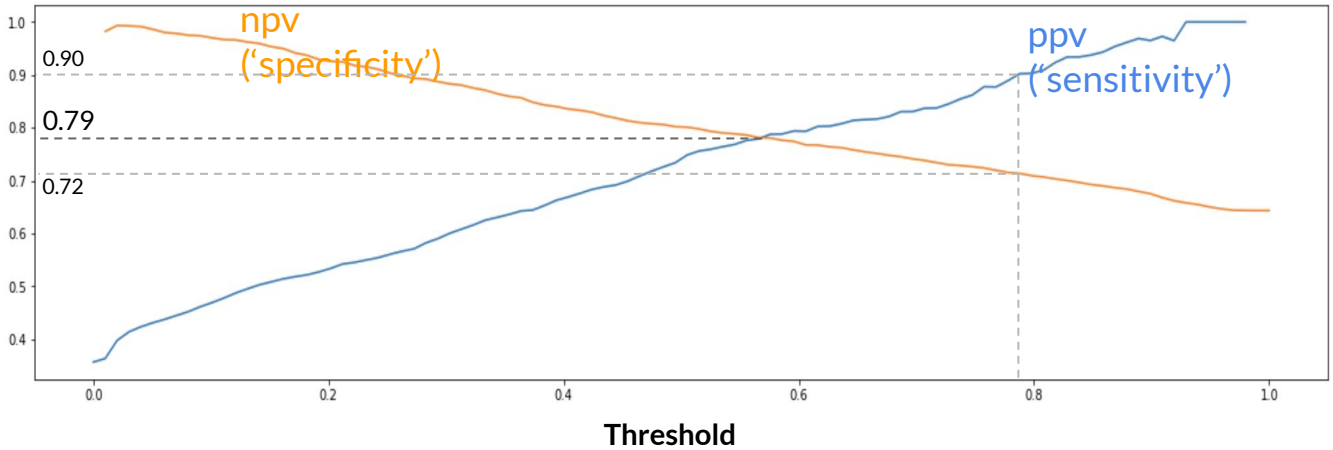
Avg. Precision: 0.805 ± 0.013

Balanced Accuracy: 0.764 ± 0.008

```
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.88      0.84      1843
           1       0.74      0.61      0.67      1022

    accuracy                           0.79      2865
   macro avg       0.77      0.75      0.76      2865
weighted avg       0.78      0.79      0.78      2865
```



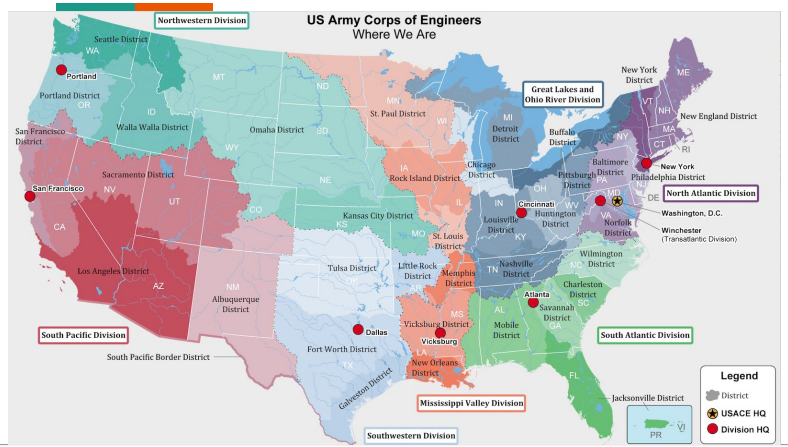
	Pred 0	Pred 1
True 0	1625	**218**
True 1	**395**	627

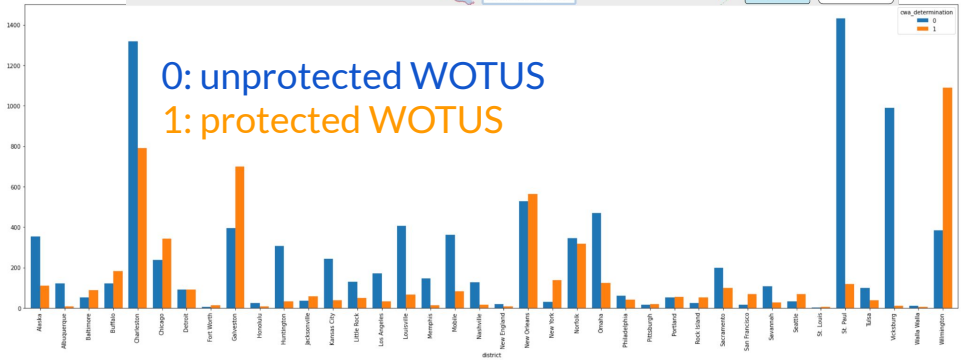
ROC, Recall, PPV, NPV by district

District	Count	%WOTUS	ROC	Recall	PPV	NPV
Honolulu	27	29	1.000	1.000	1.000	1.000
Nashville	104	10	0.513	0.200	1.000	0.882
Mobile	346	17	0.979	0.857	0.857	0.972
St. Paul	1160	7	0.953	0.762	0.800	0.983
Memphis	120	8	0.991	1.000	0.750	1.000
Rock Island	64	67	0.944	0.889	0.889	0.833
Galveston	834	63	0.914	0.920	0.865	0.839
Seattle	88	67	0.878	0.857	0.750	0.833
Alaska	359	22	0.815	0.478	0.688	0.833
New Orleans	819	49	0.810	0.805	0.752	0.738
Savannah	105	19	0.758	0.333	0.667	0.826
Wilmington	1074	71	0.774	0.965	0.749	0.708
Norfolk	510	47	0.789	0.703	0.726	0.716
Los Angeles	159	15	0.826	0.300	0.600	0.821
Little Rock	139	27	0.765	0.500	0.571	0.846
Chicago	462	58	0.730	0.785	0.739	0.622
Detroit	142	46	0.690	0.556	0.769	0.579
New York	103	74	0.730	0.952	0.909	0.500
Huntington	256	8	0.633	0.143	0.500	0.908

District	Count	%WOTUS	ROC	Recall	PPV	NPV
Philadelphia	76	38	0.626	0.429	0.600	0.733
Buffalo	220	57	0.734	0.724	0.636	0.667
Charleston	1687	36	0.691	0.557	0.507	0.743
Tulsa	109	26	0.684	0.125	0.500	0.720
Omaha	408	21	0.698	0.273	0.429	0.826
Portland	76	50	0.636	0.667	0.556	0.583
Baltimore	115	63	0.686	0.818	0.529	0.600
Sacramento	232	32	0.496	0.263	0.385	0.659
Jacksonville	66	60	0.467	0.600	0.600	0.250
Kansas City	217	13	0.708	0.125	0.143	0.857
New England	22	34	0.100	0.500	0.200	0.500
Pittsburgh	26	51	0.250	0.200	0.250	0.200
San Francisco	69	79	0.733	0.933	0.824	0.000
Louisville	371	12	0.586	0.000	0.000	0.908
Fort Worth	16	66	NaN	0.500	1.000	0.000
Vicksburg	762	1	0.657	0.000	NaN	0.982
Albuquerque	97	5	0.400	0.000	NaN	0.968
Walla Walla	13	31	1.000	0.000	NaN	0.333

Model Performance by District

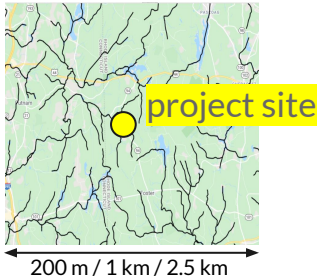




District	Count	WOTUS	ROC
Walla Walla	13	0.23	1.000
Honolulu	27	0.30	1.000
St. Louis	4	0.75	1.000
Memphis	120	0.08	0.991
Mobile	346	0.18	0.979
St. Paul	1160	0.07	0.953
Rock Island	64	0.69	0.944
Galveston	834	0.63	0.914
Seattle	88	0.70	0.878
Los Angeles	159	0.14	0.826
Alaska	359	0.22	0.815
New Orleans	819	0.48	0.810
Norfolk	510	0.47	0.789
Wilmington	1074	0.72	0.774
Little Rock	139	0.28	0.765
Savannah	105	0.18	0.758
Buffalo	220	0.59	0.734
San Francisco	69	0.78	0.733
New York	103	0.72	0.730

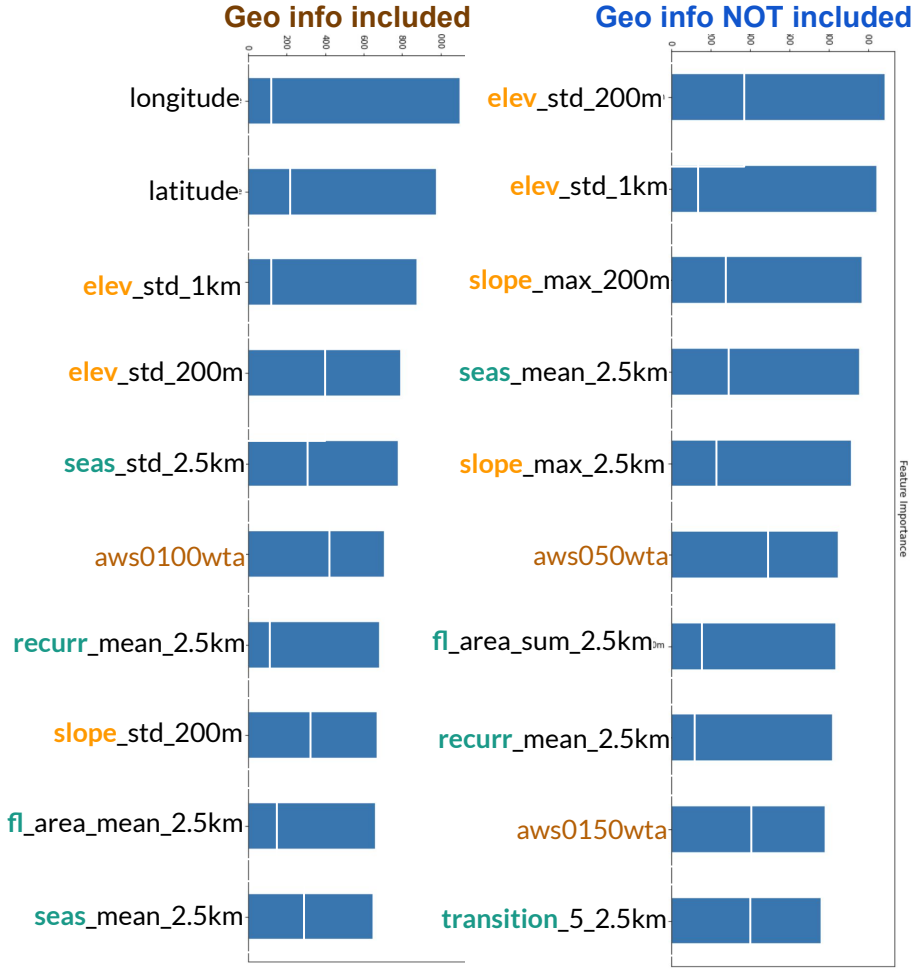
District	Count	WOTUS	ROC
Chicago	462	0.58	0.730
Kansas City	217	0.13	0.708
Omaha	408	0.21	0.698
Charleston	1687	0.37	0.691
Detroit	142	0.44	0.690
Baltimore	115	0.66	0.686
Tulsa	109	0.26	0.684
Vicksburg	762	0.01	0.657
Portland	76	0.50	0.636
Huntington	256	0.08	0.633
Philadelphia	76	0.39	0.626
Louisville	371	0.13	0.586
Nashville	104	0.09	0.513
Sacramento	232	0.32	0.496
Jacksonville	66	0.59	0.467
Albuquerque	97	0.06	0.400
Pittsburgh	26	0.50	0.250
New England	22	0.36	0.100
Fort Worth	16	0.62	NaN

Top Predictors



Geographical Information (lat, lon, district)	ROC	Top Predictors*
Included	0.87	lat, lon elevation, slope seasonality, recurrence soil extent of waterbodies
Not included	0.83	elevation, slope seasonality, recurrence soil extent of waterbodies

*200m or 1000m patch size
2.5km patch size



Model Performance vs Datasets

Datasets*	ROC_AUC
lat, lon, district (“**Geo**”)	0.841 ± 0.006
Geo + **SSURGO**	0.848 ± 0.009
Geo + **SRTM**	0.860 ± 0.005
Geo + **NWI**	0.863 ± 0.005
Geo + **NHD**	0.874 ± 0.007
Geo + **SSURGO + SRTM + NHD + NWI**	0.875 ± 0.009

*includes 200mX200m, 1kmX1km and 2.5kmX2.5km satellite image patch sizes

Sentinel 2

- 5 day re-visit period
- 10 m resolution
- Multi-spectral
- Red-edge spectral reflectance useful for vegetation detection

Google Earth

- Is a cloud computing platform with a multi-peta byte catalog of satellite imagery and geospatial datasets
- Offers APIs in JavaScript as well as in Python for making computational requests to the Earth Engine servers

Geospatial Data

- Data collected using airborne or satellite sensors
- Passive sensors measure electro-magnetic radiation or sun's energy naturally reflected from the earth's surface during the daytime.
- Types include aerial photography, multispectral & hyperspectral imagery.

Data Engineering

Sentinel-2 Remote Sensing

Satellite Indices

$$\text{NDVI} = \frac{(\text{NIR} - \text{Red})}{(\text{NIR} + \text{Red})}$$

$$\text{NDWI} = \frac{(Xnir - Xswir)}{(Xnir + Xswir)}$$

MNDWI = (Green − SWIR) / (Green + SWIR)

$$MNDVI = NDVI \times (SWIR_{max} - SWIR) / (SWIR_{max} - SWIR_{min})$$

NWI (National Wetland Inventory)

- Initiated in 1974; produced by manually interpreting mid-1980's aerial photographs at a scale of 1:24K supported by soil surveys and field verifications.
- Presently covers mapping for 90% of United States (30% Alaska)
- Made available for digital downloads in 2016 as ESRI GeoDatabase or Shapefile.
- Mapping units vary between 1000 and 20,000 sq.metres based on type of aerial imagery used originally.
- Considered most accurate for permanently flooded wetlands
- Considered conservatively accurate for seasonal, temporarily flooded, ephemeral and forested wetlands
- **Does not capture area changes over past 30 years that are due to natural and human activities.**

NHD (National Hydrography Dataset)

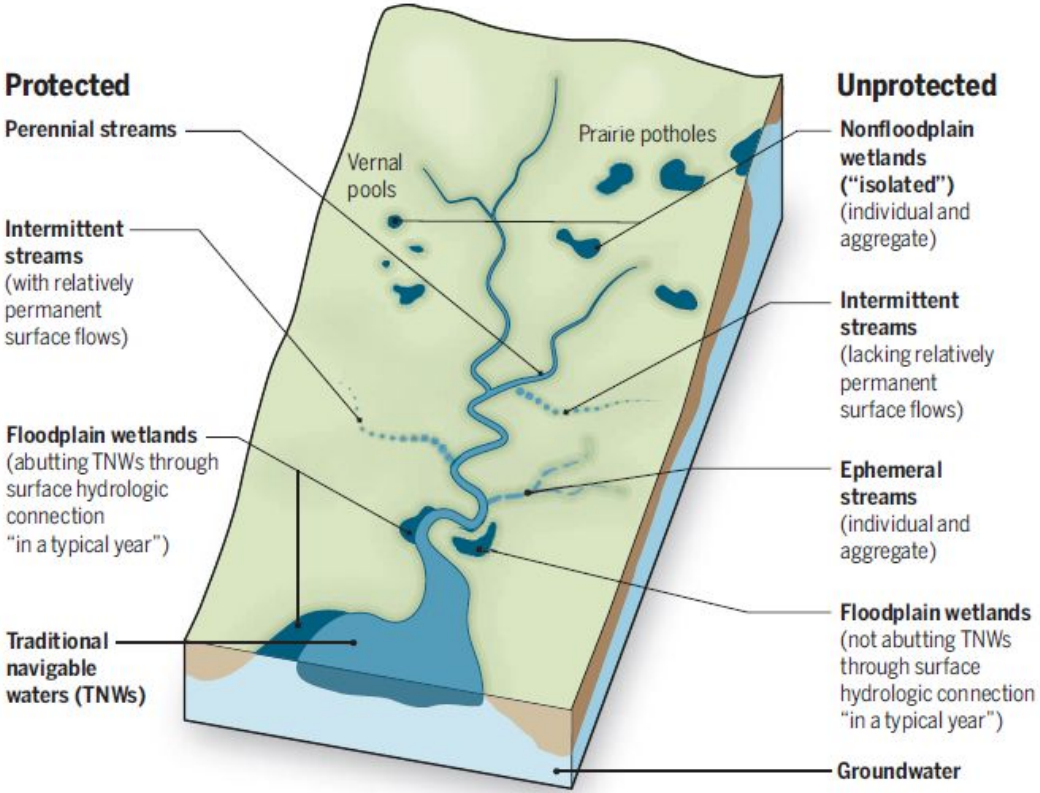
- The National Hydrography Dataset (NHD) represents the water drainage network of the United States with features such as rivers, streams, canals, lakes, ponds, coastline, dams, and stream gauges.
- The NHD is the most up-to-date and comprehensive hydrography dataset for the Nation.
- The U.S. EPA along with the USGS was responsible for its creation and maintenance
- In the late 1990s, the USGS and the US EPA collaborated to produce the medium resolution National Hydrography Dataset at 1:100,000 scale for the conterminous U.S. - which is the version we are using for reason of completeness.

SSURGO (Soil Survey Geographic Database)

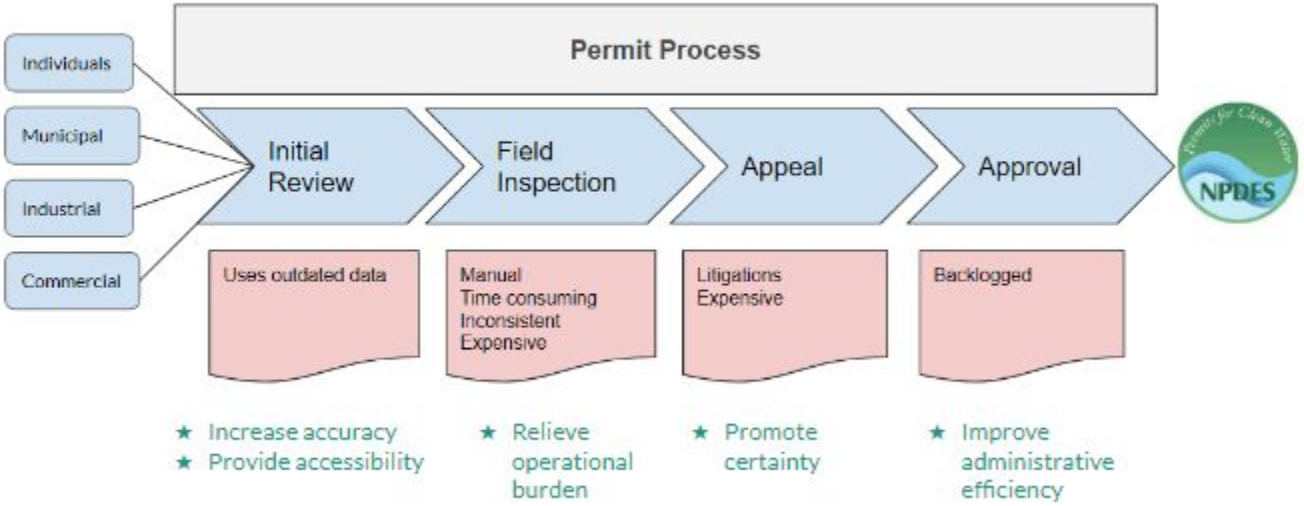
- The SSURGO database contains information about soil as collected by the National Cooperative Soil Survey over the course of a century.
- Many soil samples were analyzed in laboratories.

Protected versus unprotected waters

Multiple waterbody types were initially under consideration for protection as "waters of the United States" under the Navigable Waters Protection Rule. Ephemeral streams flow only after precipitation events, intermittent streams flow periodically or seasonally, and perennial streams flow continuously. There are many types of nonfloodplain, or "isolated" wetlands, including prairie potholes and vernal pools, as illustrated here.



How can this technique help?



- Platform for identifying jurisdictional waters more easily and quickly
- Maps to facilitate understanding of CWA jurisdiction
  - promote greater regulatory certainty,
  - relieve some of the regulatory burden associated with determining the need for a permit