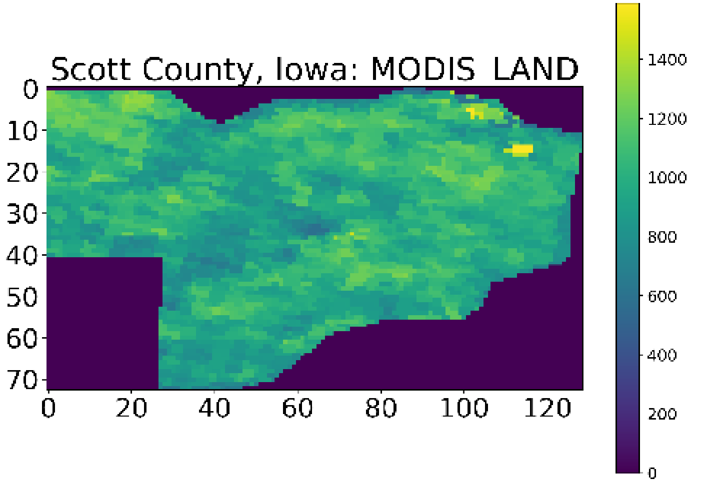
Early Prediction of Corn Yields Across U.S. from Satellite Images

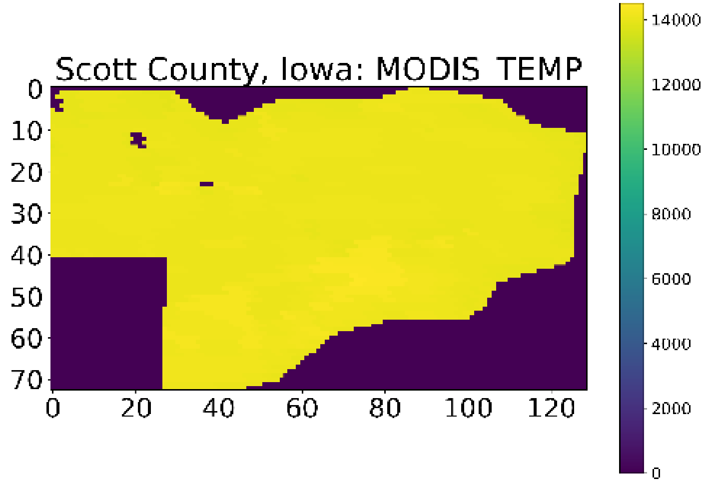
Introduction

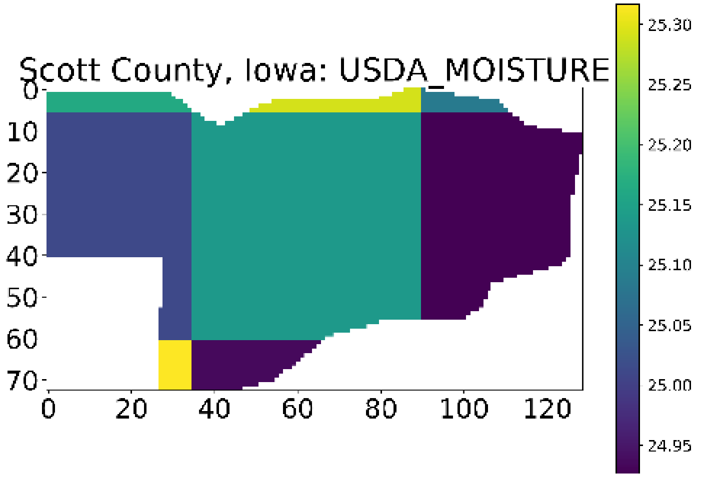
- **Motivation**
  - A way to forecast food shortage, food supply, and hence better price and inventory management.
  - Benefits agricultural distributors, food processing industries, and government

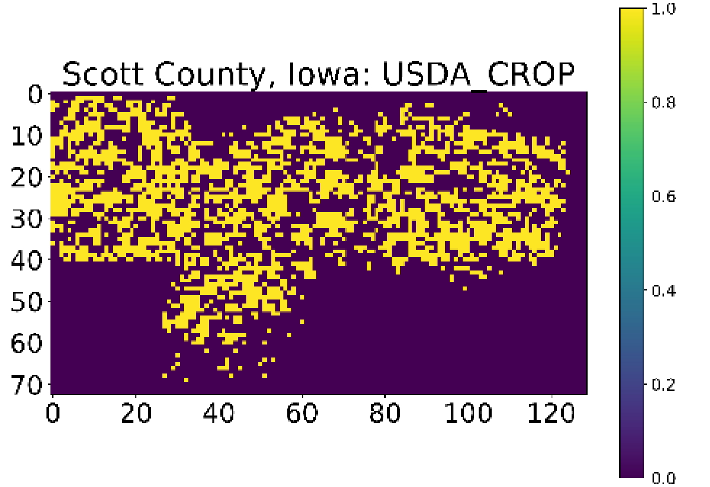
- **Dataset (total of 146 GB)**
  - MODIS Terra Surface Reflectance
  - MODIS Surface Temperature
  - USDA-FAS Surface and Subsurface Moisture
  - USDA-NASS for masking
  - The ground truth annual yields were collected in a county-level from USDA QuickStats.

Quick Look at the Data: Image Sample



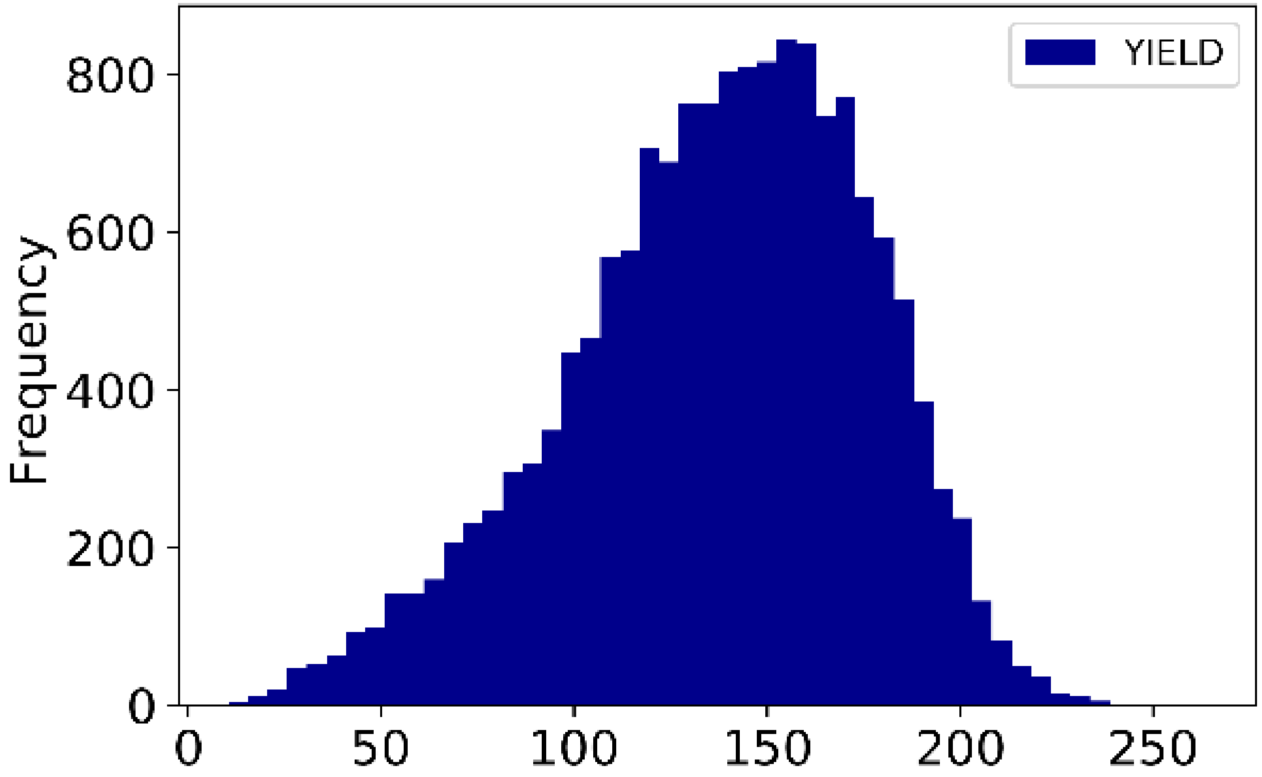






- MODIS Terra Surface Reflectance: 11 bands
- MODIS Surface Temperature: 2 bands
- USDA-FAS Surface and Subsurface Moisture: 2bands
- USDA-NASS for masking: 1 band

Quick Look at the Data: Ground Truth

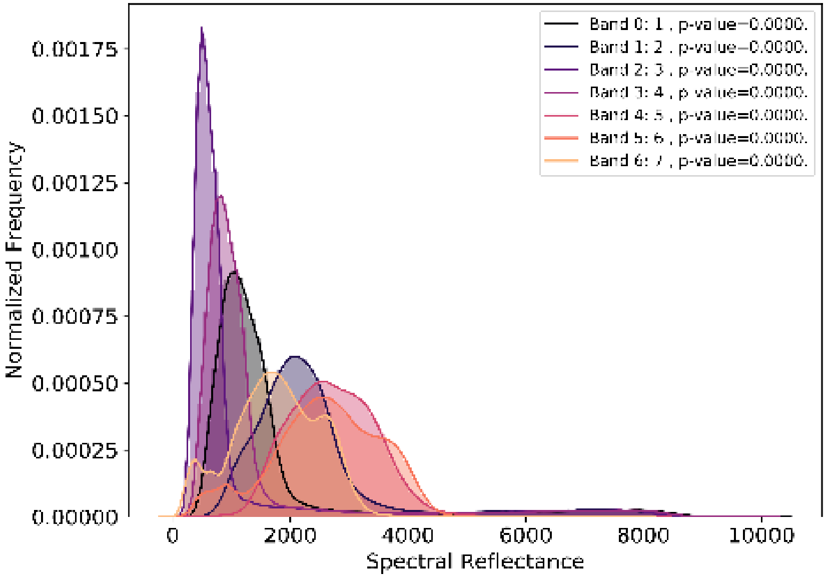


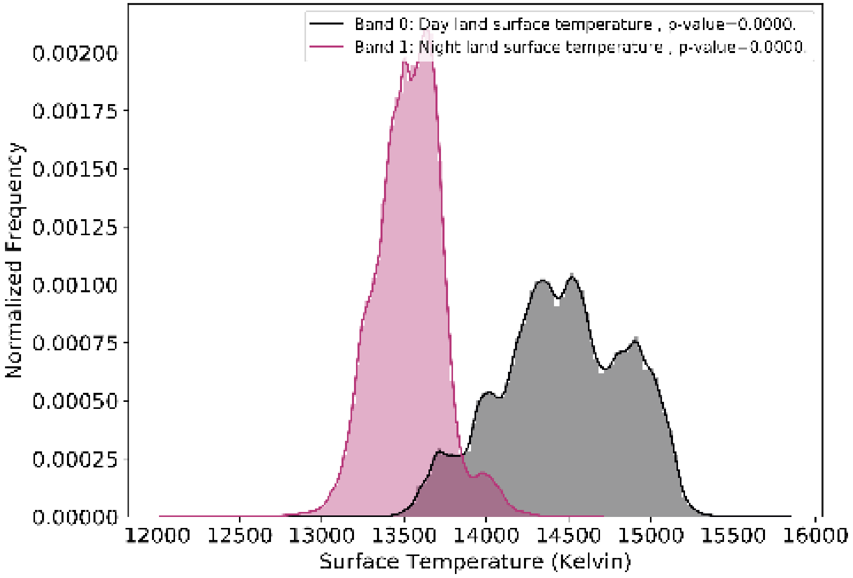
- Yield distribution of all counties from year 2010 to year 2016
- The yield is not normally distributed

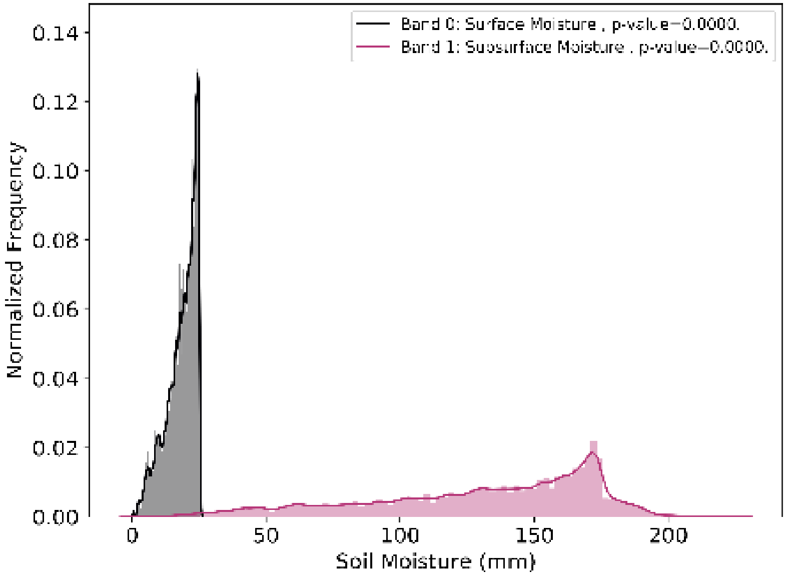
Image Preprocessing Steps

1. Fill NaN values with zeros
2. Combining images that have higher cadences using its mode values
3. Separate set of images to each year
4. Stack image bands from different satellites
5. Mask all bands of images to 0 except pixels with corn fields
6. Delete images that only contain zeros after masking

Image Value Distribution After Masking

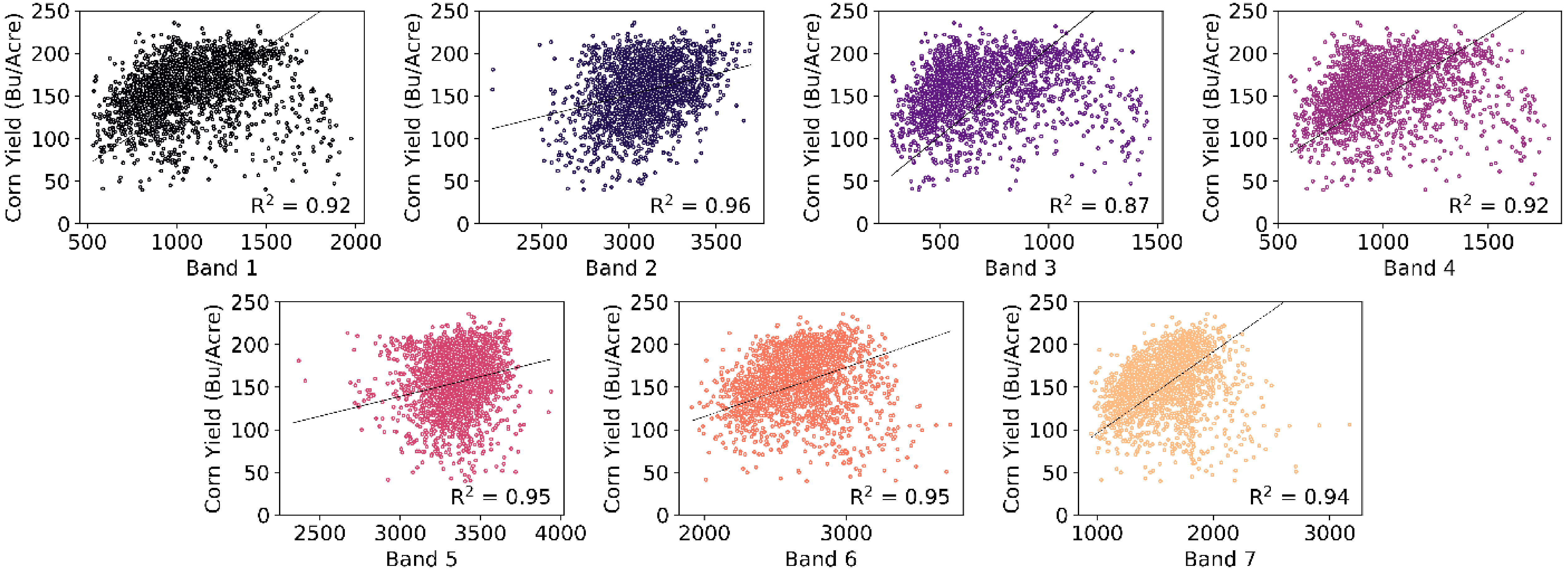




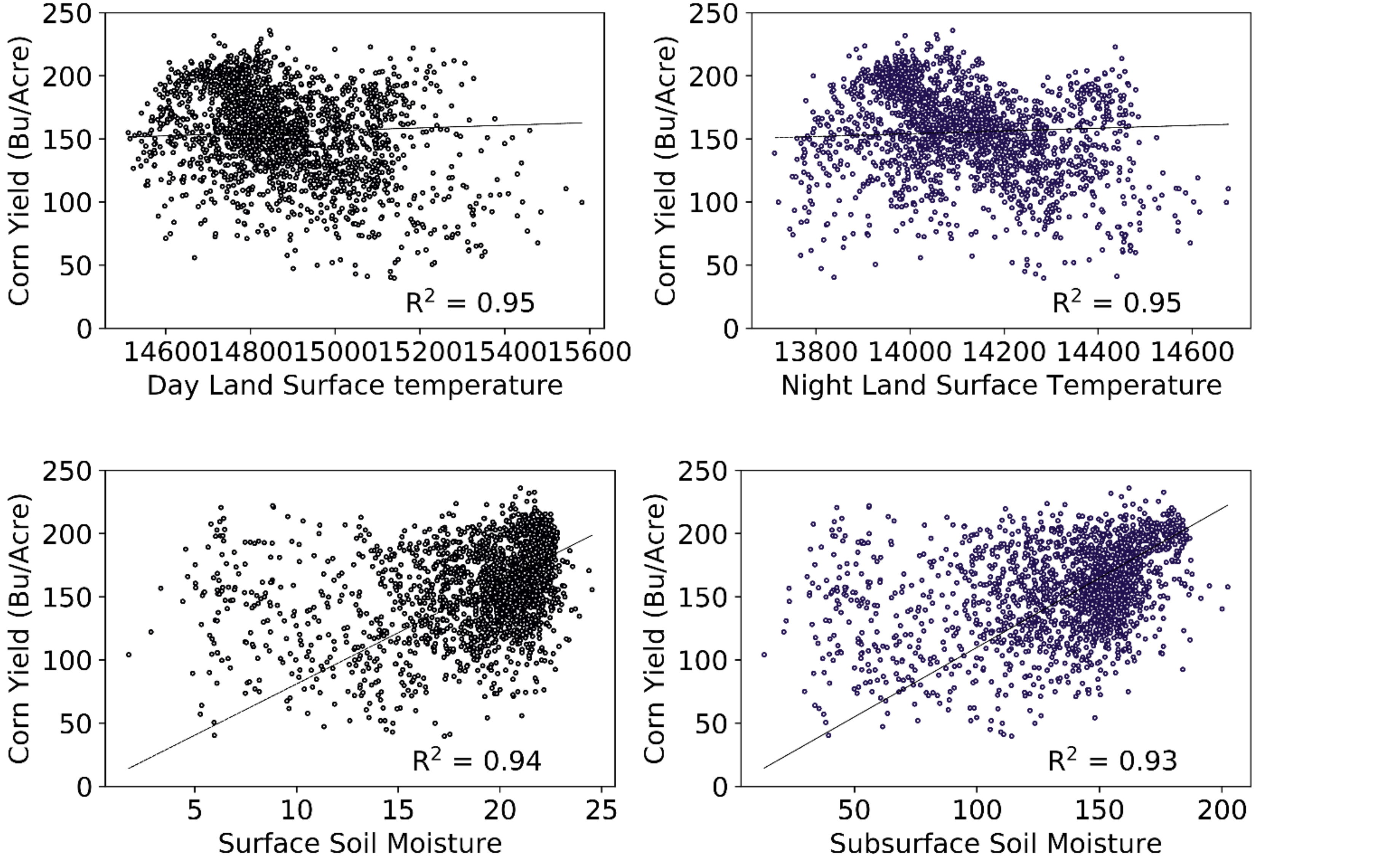


- These value distribution will be important for feature engineering later in the study

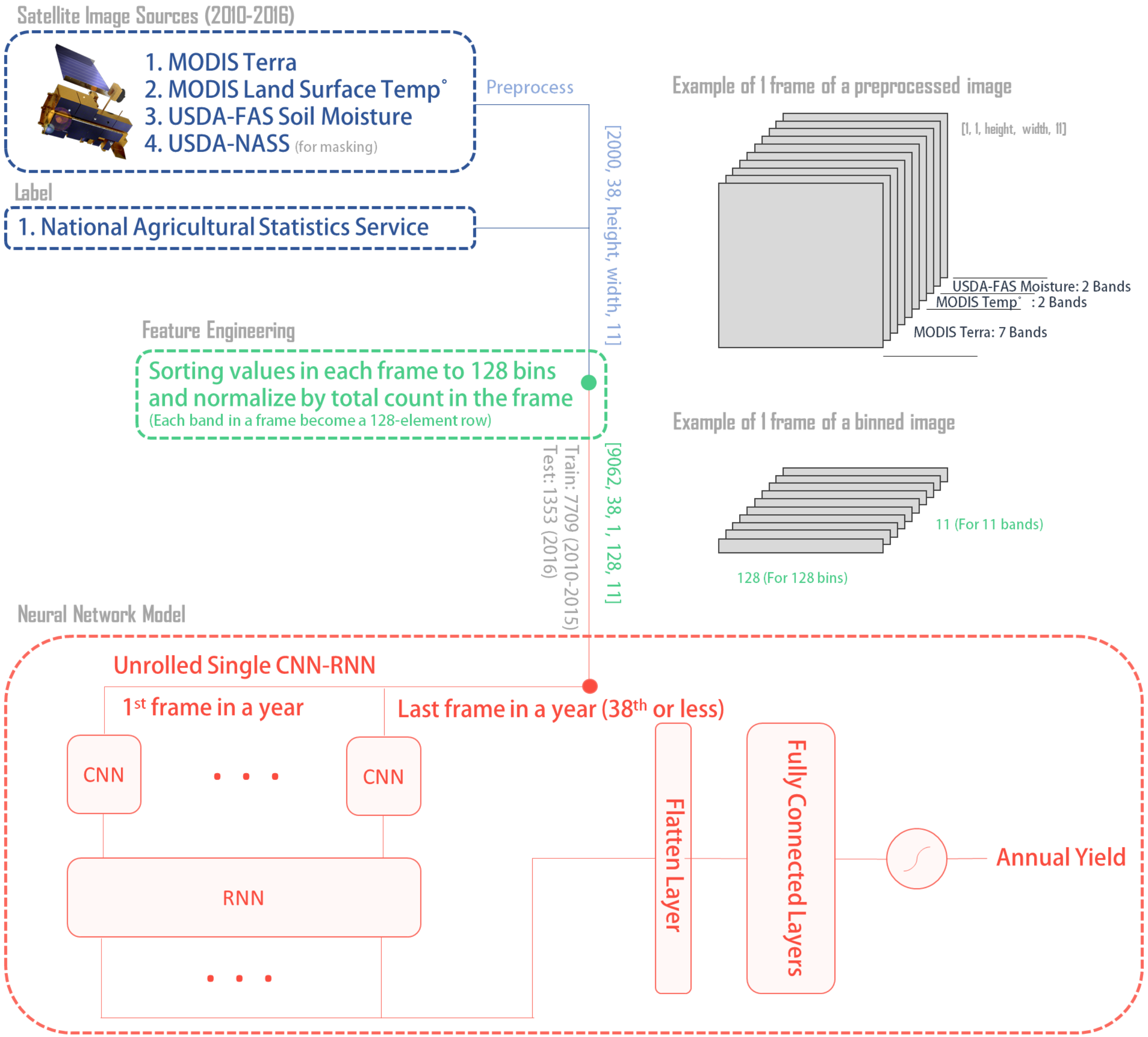
Exploratory Data Analysis: MODIS Terra Reflectance



Exploratory Data Analysis: MODIS Temp & USDA Moisture



Workflow Summary: from Data to Model



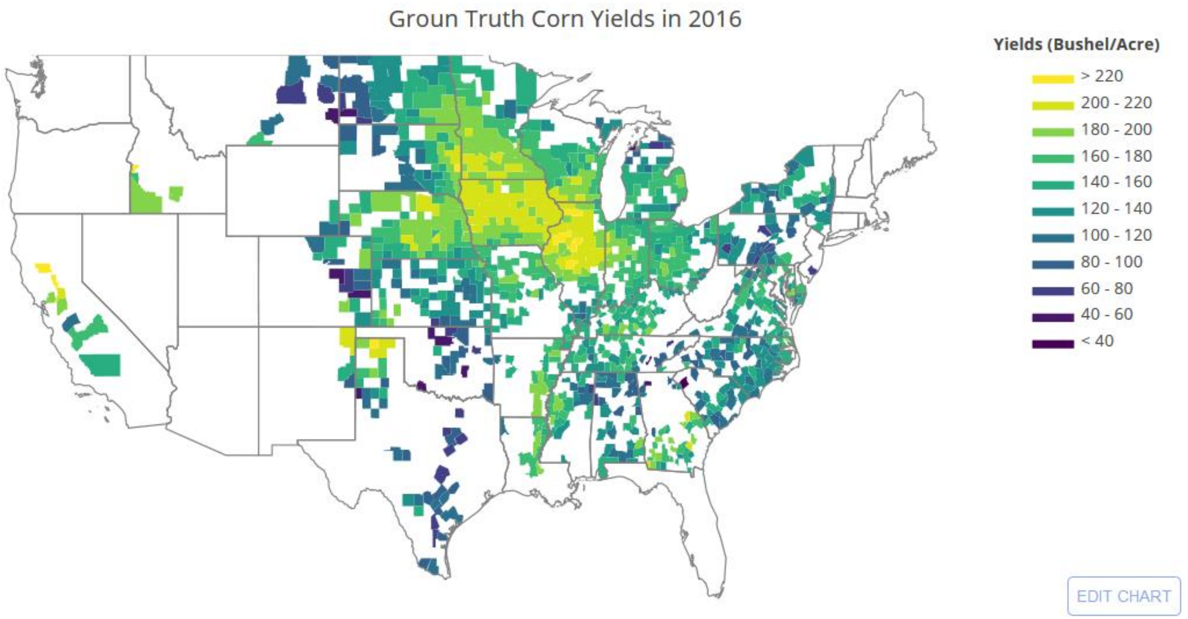
- Using raw images for prediction would be extremely expensive because each image is about 100 x 100 x 11 and each year there are 38 images.
- After all preprocessing we have a total of 9000 images left (from year 2010-2016).
- Data from year 2016 will be used for test set, the rest is for training.

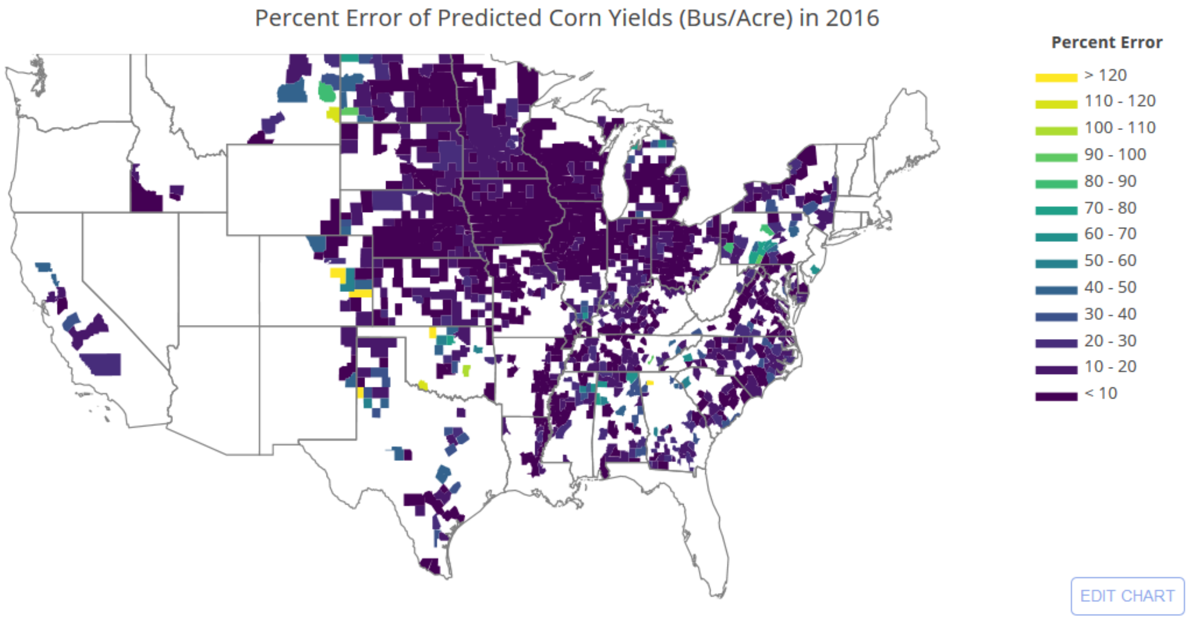
Model Results and Batch Size Optimization

Model	Mean Absolute Error (Bushel/Acre)	Percent Error from Mean
Custom CNN-LSTM	18.85	12.17%
Separable CNN-LSTM	24.11	15.57%
CNN-LSTM [5]	**16.20**	**10.46%**
3D-CNN	96.39	62.26%
CNN-LSTM-3D CNN	36.14	23.34%

Batch Size	Mean Absolute Error (Bushel/Acre)
8	61.92
16	17.17
32	**16.20**
64	20.10

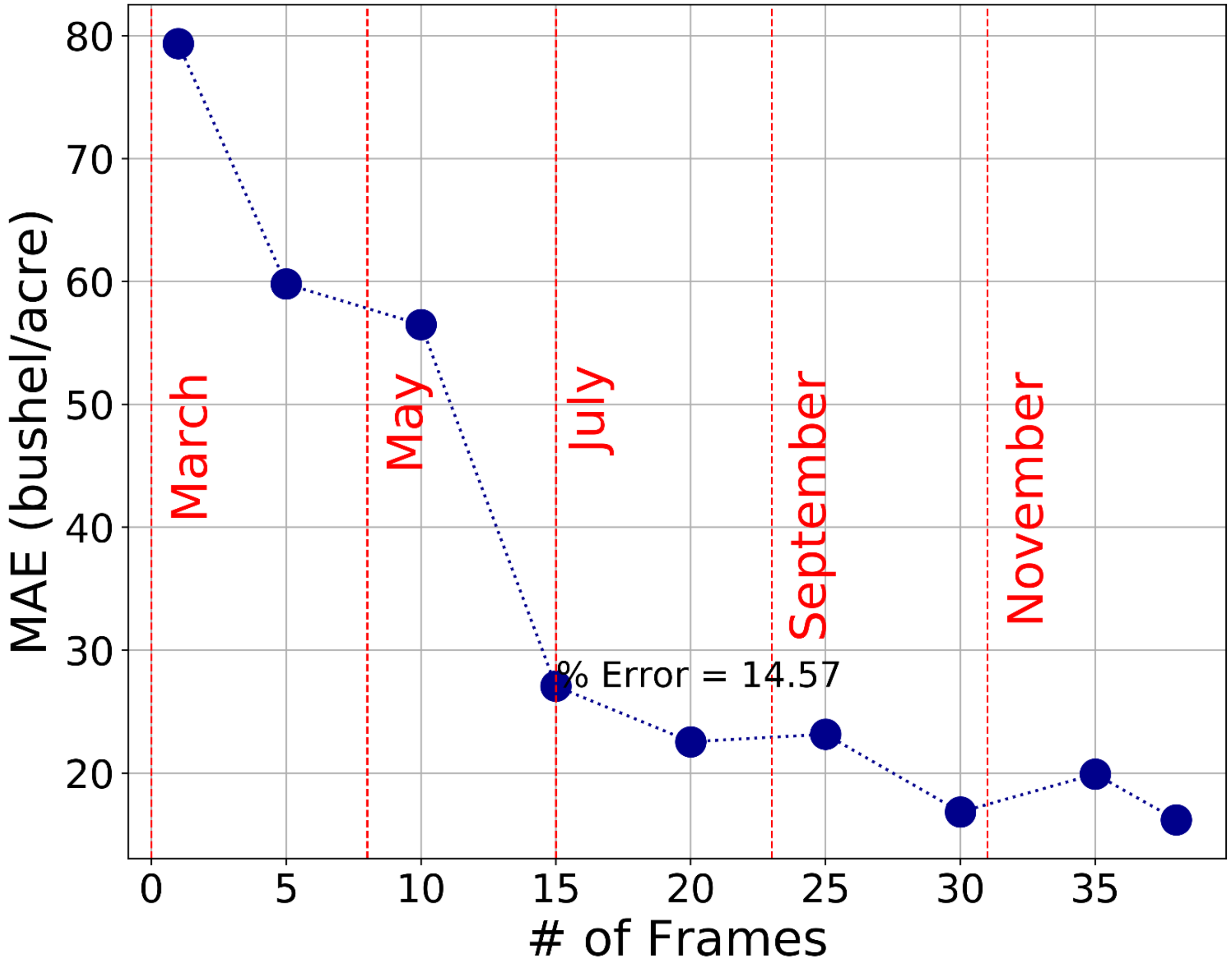
Result Analysis: Geospatial Analysis





- Model tends to perform poorer in area with extremely low yields such as some parts of Montana and North Dakota.
- This could be due to the low amount of samples with this extreme.
- On the other hand, the model performs extremely well (percent error <10%) in the case with typical yields to high yields such as those in Iowa, Nebraska, and Illinois.

Result Analysis: Reducing the Number of Frames



- Using just 20 frames (roughly 2nd week of Aug.), we can already achieve percent error as low as 14.57% (compared to 10.46% if we use the images from the entire year).
- This is about 2 months before corn is typically harvested in October, although this could even be later in the year in warmer states.

Conclusion

- we have shown the correlation between different satellite images including reflectance, land temperature, and land moisture to corn yields in U.S.
- We leveraged these correlations to construct a CNN+RNN model that can capture both spatial and temporal information of these data to predict corn yields in a year.
- The model was able to predict with MAE of 10% at 38 frames per year, and as low as 15% even if only 20 frames were used. This is equivalent to taking images upto just August.
- Therefore, this model would allow user to be able to predict corn yields at county level early in the season.