ATMS 597 Project 6 Group B

Carolina Bieri, Chu-Chun Chen, and Jeffrey Thayer

Our Task

Can information about the MJO and/or soil moisture (i.e. slowly-evolving factors) be used to predict meteorological variables in S. America?

1. **Regression** - Predicting exact numbers
    a. Multiple Linear Regression (baseline)
    b. Random Forest
    c. Neural Network

2. **Classification** - Predicting whether a threshold is exceeded
    a. Logistic Regression (baseline)
    b. Neural Network

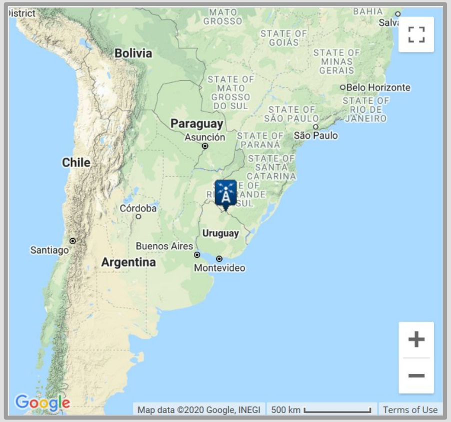
Can information about the MJO and/or soil moisture (i.e. slowly-evolving factors) be used to predict meteorological variables in S. America?

Data Sources:

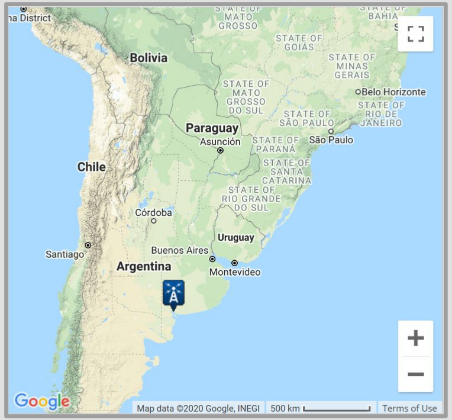
- -MJO index (OLR+U850+U200)
- -ERA5 soil moisture (2 layers)
- -GHCN station data
  - -TMAX, TMIN, PRCP

Daily data for 1980-2019

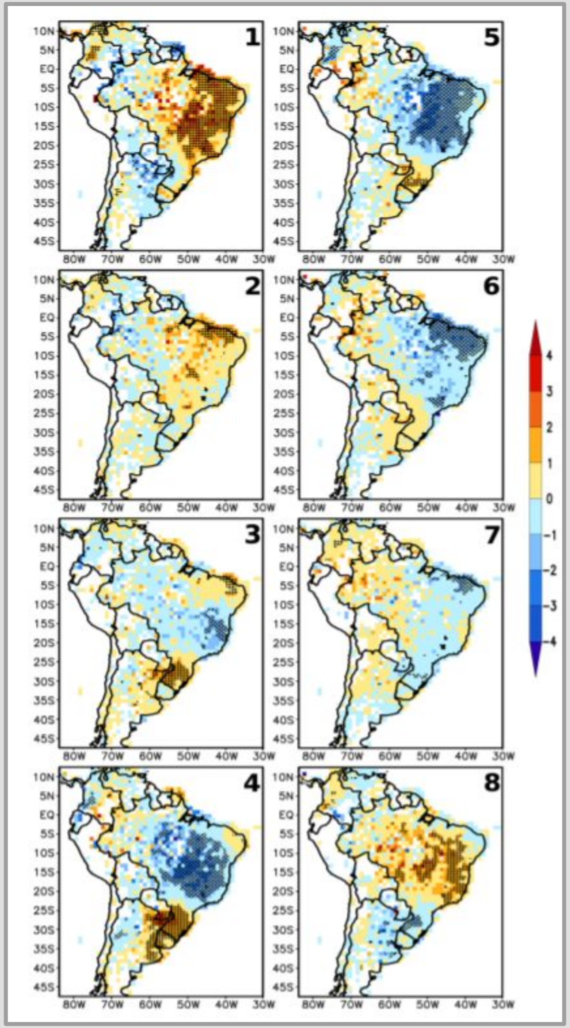
Uruguay station



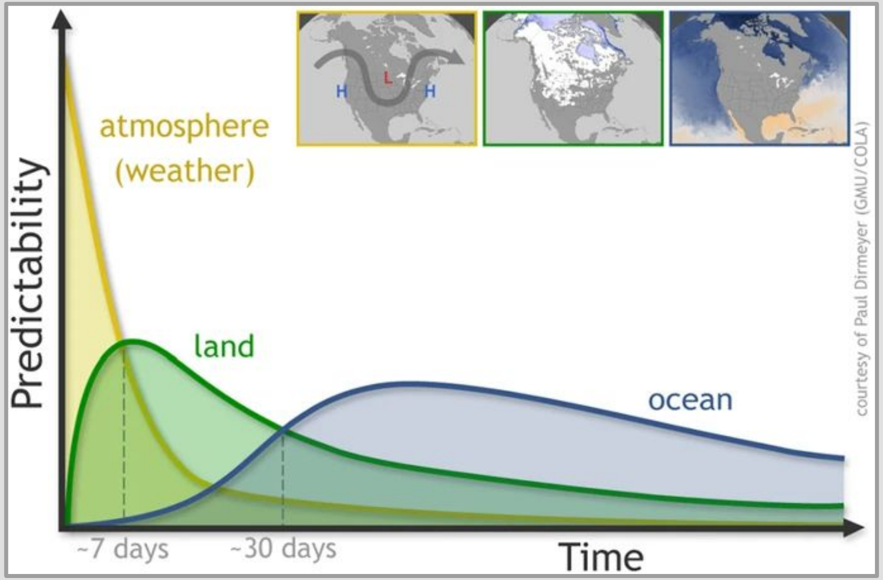
Argentina station



Grimm (2019):
Influence of MJO phase on precipitation anomalies during austral summer



Mariotti et al. (2018):
Sources of predictability for the atmosphere



Data Processing

Data Processing steps

1. Combine Station, MJO, and Soil Moisture data into single dataframe
   a. 1 dataframe per station
2. Add lagged columns (5-30 days at 5-day intervals) for MJO and Soil Moisture
3. Drop NaN rows after removing unneeded columns
   a. Using all months
4. Divide into training and testing sets (70-30)
5. Binarize (if classifying)
6. Normalize based on training data if necessary
   a. Neural network - Standardized anomalies

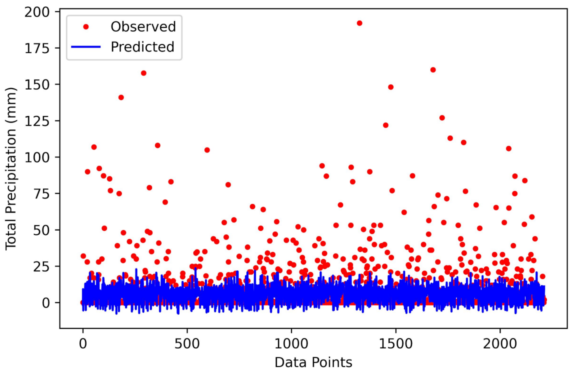
Results

Multiple Linear Regression

Total Precipitation

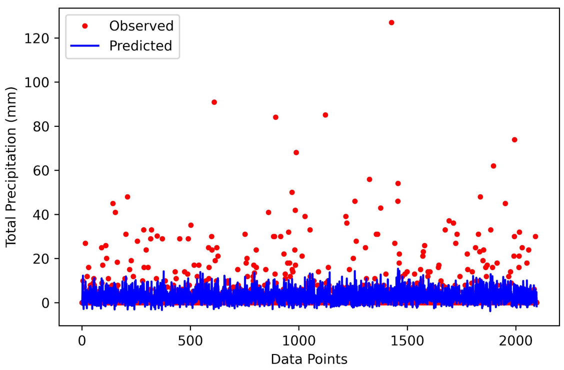
Uruguay RMSE: **15.45 mm**

- 7371 data points
- Lag0 layer 1 Soil moisture
  Lag5 layer 1 Soil moisture



Argentina RMSE: **7.59 mm**

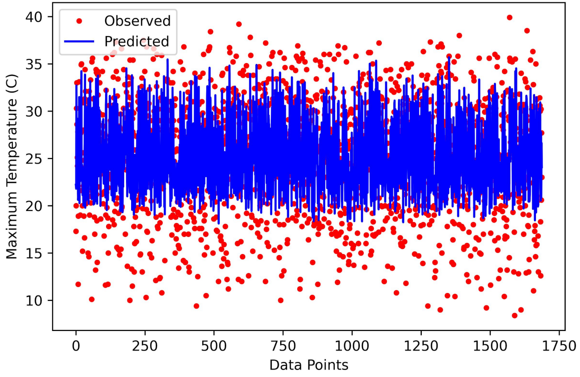
- 6990 data points
- Lag0 layer 1 Soil moisture
  Lag0 layer 2 Soil moisture



Maximum Temperature

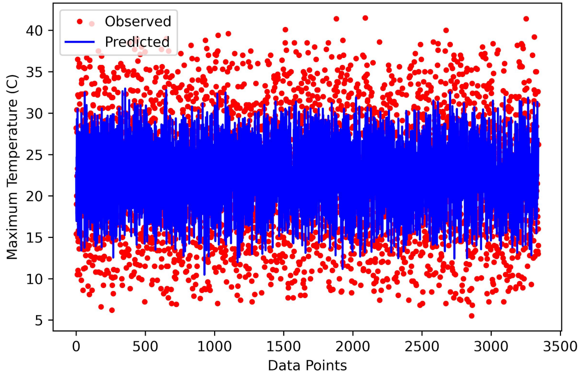
Uruguay RMSE: **4.90°C**

- 5625 data points
- Lag0 layer 2 Soil moisture
  Lag5 layer 2 Soil moisture



Argentina RMSE: **6.12°C**

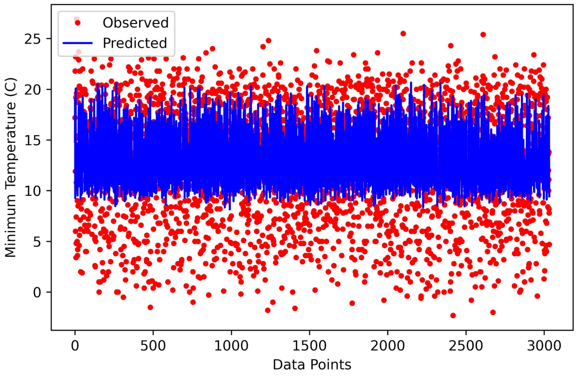
- 11141 data points
- Lag0 layer 1 Soil moisture
  Lag30 layer 2 Soil moisture



Minimum Temperature

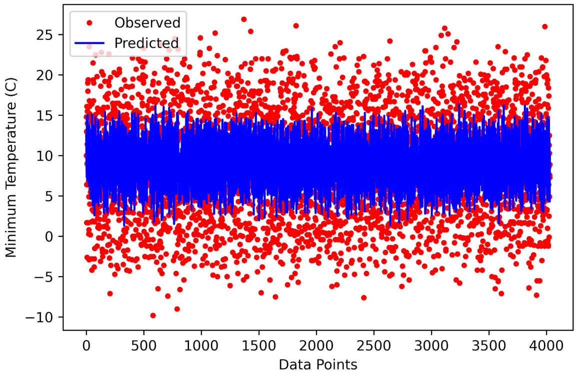
Uruguay RMSE: **4.84°C**

- 10113 data points
- Lag0 layer 2 Soil moisture
  Lag5 layer 1 Soil moisture
  Lag0 MJO amplitude



Argentina RMSE: **5.73°C**

- 13431 data points
- Lag0 layer 2 Soil moisture
  Lag30 layer 2 Soil moisture



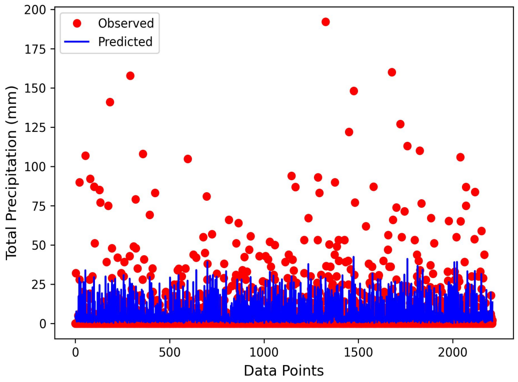
Random Forest

- Hyperparameter optimization using RandomizedSearchCV for 300 random combinations
  - N_estimators = **800**, random_state = 42, max_depth = **100**, max_features = **'sqrt'**, Min_samples_leaf = **4**, min_samples_split = **10**, bootstrap = True

Total Precipitation

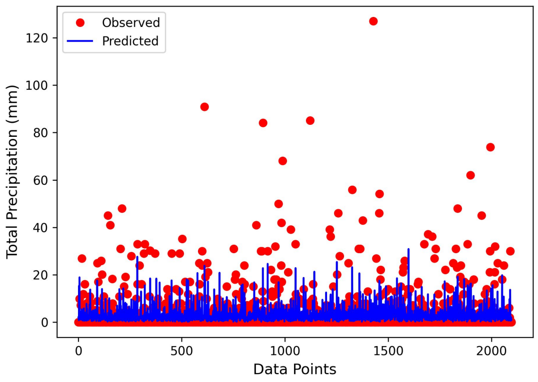
Uruguay RMSE: **14.00 mm**

- 7371 data points
- Lag0 Soil moisture and Lagged MJO amplitude



Argentina RMSE: **7.33 mm**

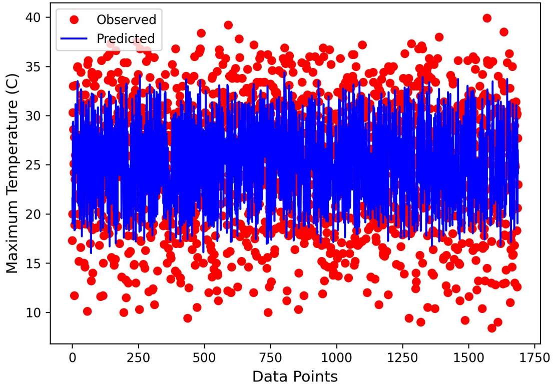
- 6990 data points
- Lag0 Soil moisture and Lagged MJO amplitude



Maximum Temperature

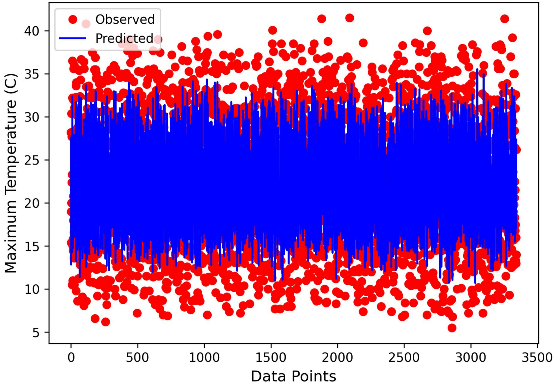
Uruguay RMSE: **3.83°C**

- 5625 data points
- Month and Lag0 Soil moisture



Argentina RMSE: **3.90°C**

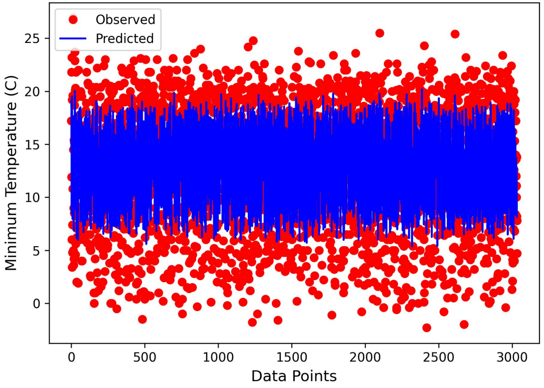
- 11141 data points
- Month and Lag0 Soil moisture



Minimum Temperature

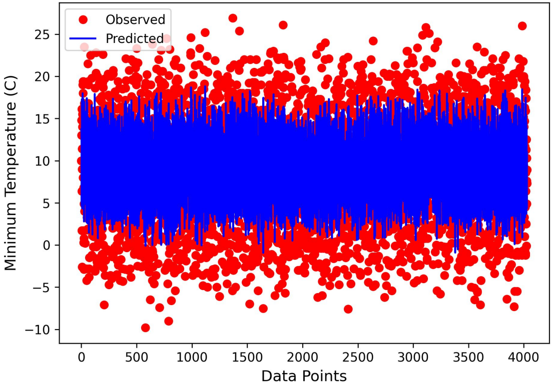
Uruguay RMSE: **3.49°C**

- 10113 data points
- Month and Lag0 Soil moisture



Argentina RMSE: **3.99°C**

- 13431 data points
- Month and Lag0 Soil moisture



Neural Network

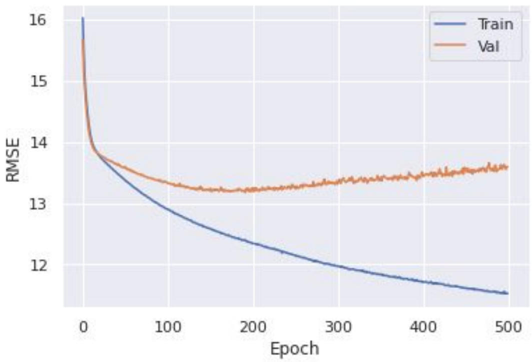
Neural Network setup

- Small, simple model
  - One layer with 10/20 units
  - 500 epochs
  - Larger capacity results in extreme overfitting
  - Need more data?
- Loss functions
  - Regression: MSE
  - Classification: Categorical cross entropy

Total Precipitation (Regression)

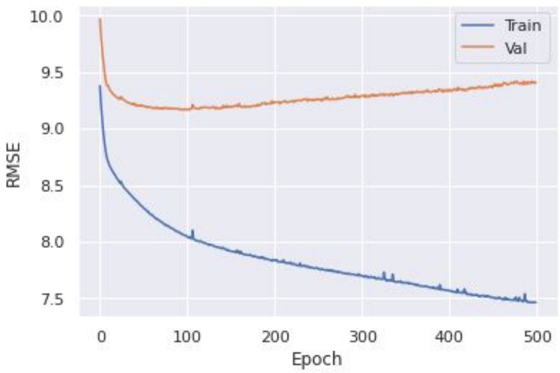
Uruguay RMSE: **14.77 mm**

- 7371 data points



Argentina RMSE: **20.53 mm**

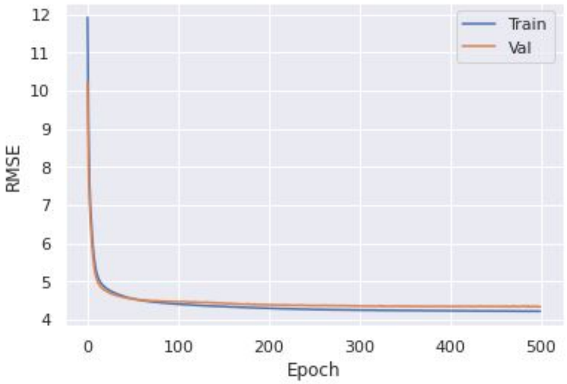
- 6990 data points



Maximum Temperature (Regression)

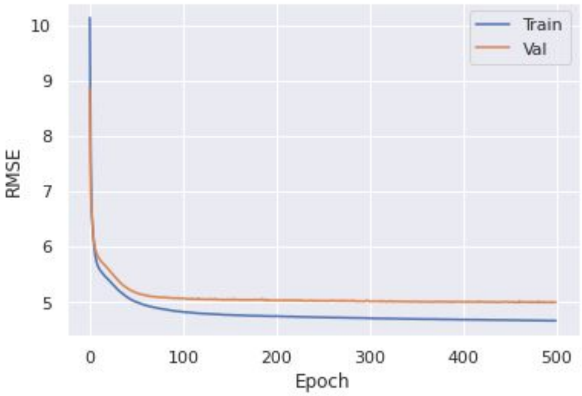
Uruguay RMSE: **4.47 °C**

- 5625 data points



Argentina RMSE: **7.01 °C**

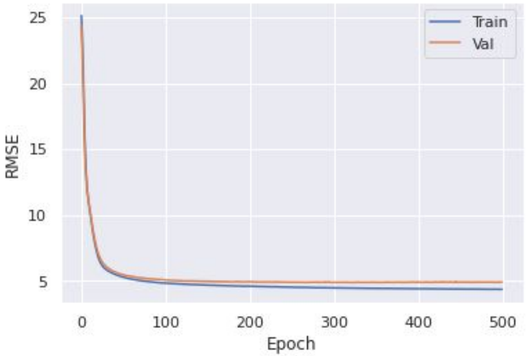
- 11141 data points



Minimum Temperature (Regression)

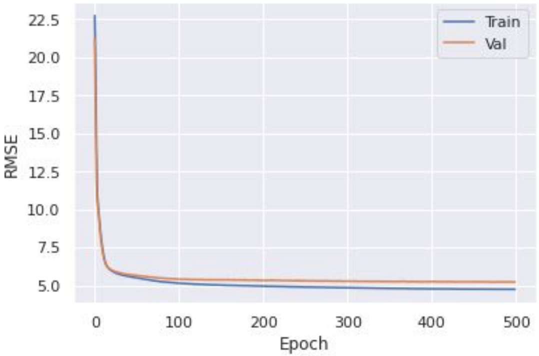
Uruguay RMSE: **4.7 ºC**

- 10113 data points



Argentina RMSE: **6.9 ºC**

- 13431 data points



Classification task

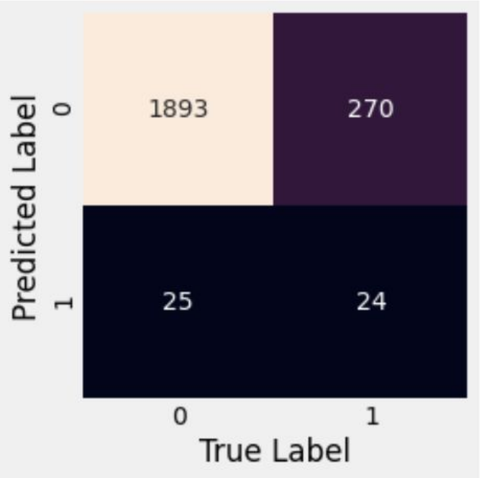
- Model might not be doing well for precipitation since our predictors act to modulate instead of having direct influence
- Try a classification task instead
    - Predict if precipitation is above 0.75 standard deviations or not
        - 0 = no, 1 = yes

Logistic Regression

Total Precipitation

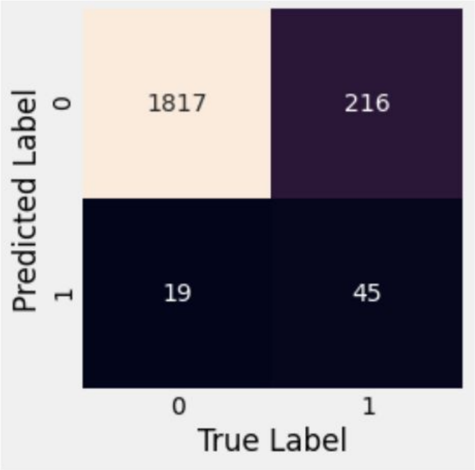
Uruguay

- Brier Skill Score: **0.15**
- 7371 data points
- Lag0 Soil moisture layer 1 and layer 2



Argentina

- Brier Skill Score: **0.20**
- 6990 data points
- Lag0 Soil moisture layer 1 and layer 2

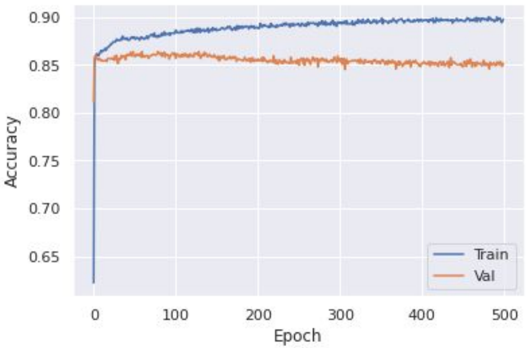


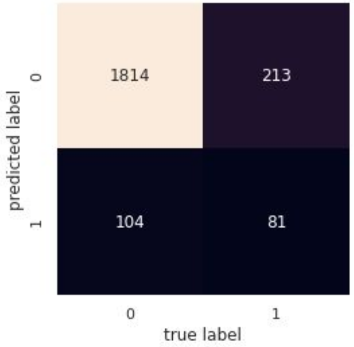
Neural Network

Total Precipitation (Classification)

Uruguay BSS: **0.18**

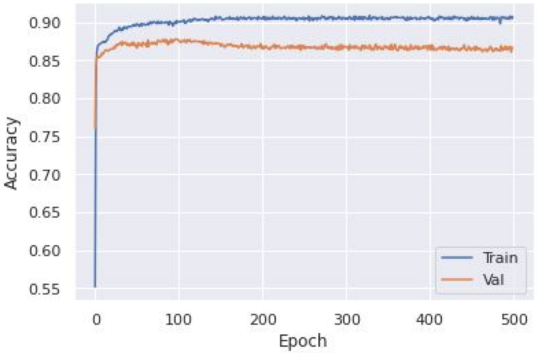
- 7371 data points
- Accuracy: 0.85

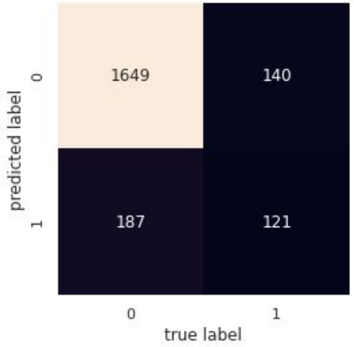




Argentina BSS: **-0.67**

- 6990 data points
- Accuracy: 0.83





Conclusion

- Help us
- Maybe add more features
- More tuning of NN
- More data?

Updates:
using GPCP precipitation in SESA (25S-30S, 50W-55W)

Can information about the MJO and/or soil moisture (i.e. slowly-evolving factors) be used to predict meteorological variables in S. America?

Data Sources:

-MJO index (OLR+U850+U200)

-ERA5 soil moisture (2 layers)

- averaged over 25-30S, 50-55W

-GPCP daily precipitation data

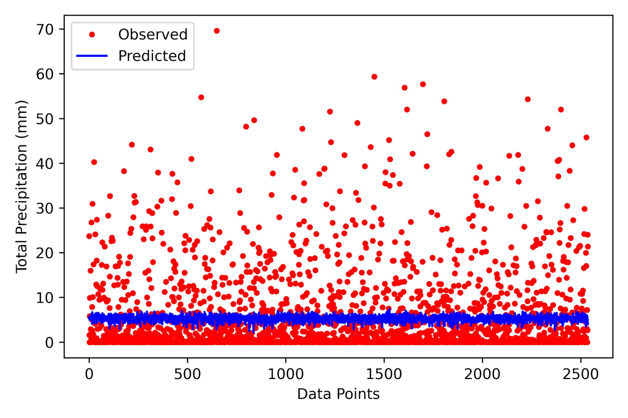
- averaged over 25-30S, 50-55W

Daily data for 1996-2019

Data Processing steps

1. Combine GPCP, MJO, and Soil Moisture data into single dataframe
2. Add lagged columns (5-30 days at 5-day intervals) for MJO and Soil Moisture
3. Drop NaN rows after removing unneeded columns
   a. Using all months
4. Divide into training and testing sets (70-30)
5. Binarize (if classifying)
6. Normalize based on training data if necessary
   a. Neural network - Standardized anomalies

Multiple Linear Regression



Multiple Linear Regression RMSEs:

All Vars, All Months, All MJO Phases: **9.30mm**
All Vars, Oct-Mar, All MJO Phases: **9.64mm**

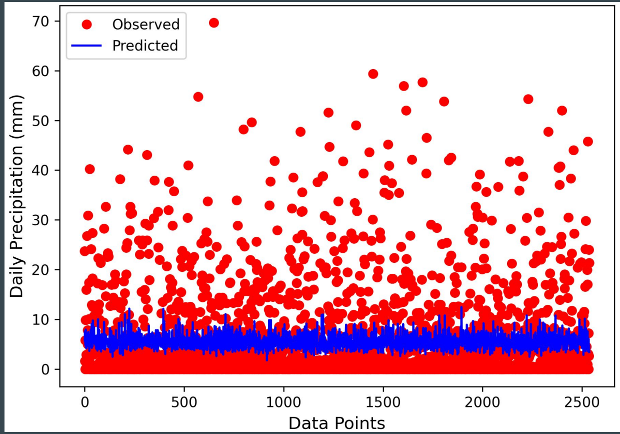
No Soil Moisture Vars, All Months, All MJO Phases: **9.30mm**
No Soil Moisture Vars, Oct-Mar, All MJO Phases: **9.64mm**

No MJO Vars, All Months: **9.30mm**
No MJO Vars, Oct-Mar: **9.63mm**

Summary:

- When using all months, almost no difference between using different sets of variables
- When using Oct-Mar, using only soil moisture variables shows a really small improvement
- Limiting months seems to hurt the prediction

Random Forest Regression



Random Forest Regression RMSEs:

All Vars, All Months, All MJO Phases: **9.39mm**
All Vars, Oct-Mar, All MJO Phases: **9.69mm**

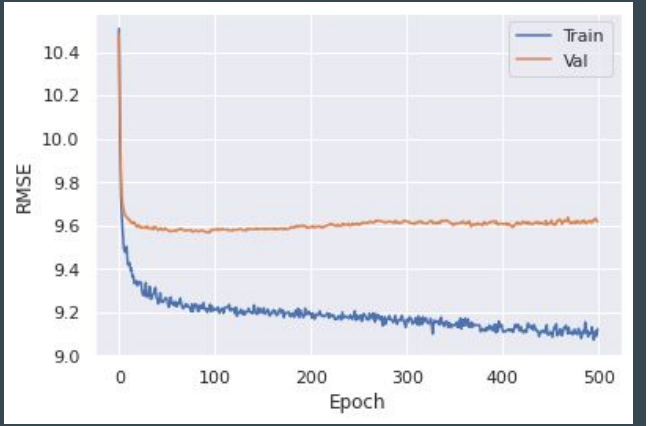
No Soil Moisture Vars, All Months, All MJO Phases: **9.40mm**
No Soil Moisture Vars, Oct-Mar, All MJO Phases: **9.78mm**

No MJO Vars, All Months: **9.39mm**
No MJO Vars, Oct-Mar: **9.67mm**

Summary:

- When using all months, almost no difference between using different sets of variables
- For only Oct-Mar months, using only soil moisture variables shows a really small improvement
- Limiting months, or limiting MJO phases (not shown), seems to hurt the prediction

Neural Network Regression

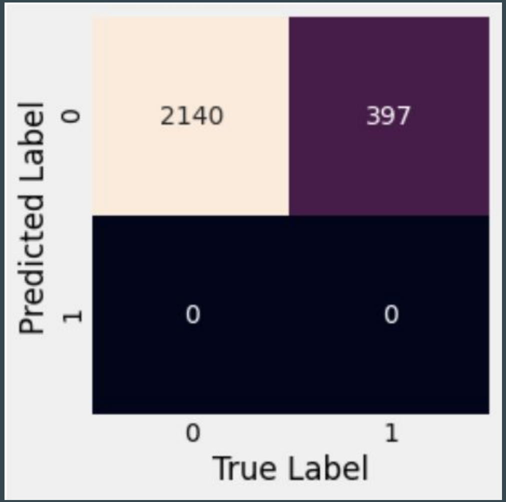


NN Regression RMSE:

All Vars, All Months, All MJO Phases: **9.36 mm**

- Use of GPCP spatial average as target results in somewhat improved predictions
- Layers: 1 (deep network gives very large RMSE)
- Neurons per layer: 16
- Learning rate: 0.001
- Epochs: 500
- Optimizer: RMSprop
- Loss function: MSE
- L2 regularization and dropout applied

Logistic Regression



Logistic Regression BSS:

All Vars, All Months, All MJO Phases: **0.00**
All Vars, Oct-Mar, All MJO Phases: **-0.01**

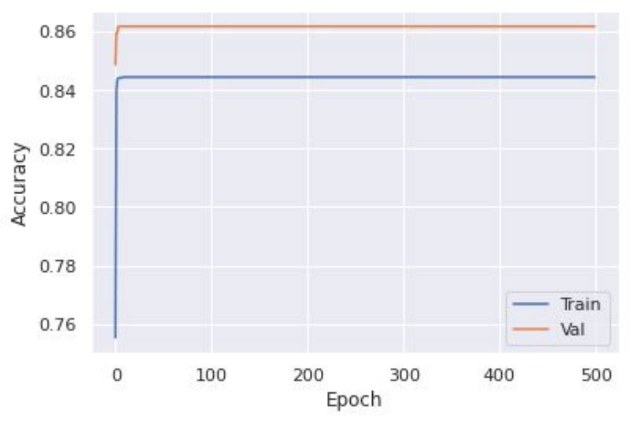
No Soil Moisture Vars, All Months, All MJO Phases: **0.00**
No Soil Moisture Vars, Oct-Mar, All MJO Phases: **-0.01**

No MJO Vars, All Months: **0.00**
No MJO Vars, Oct-Mar: **0.00**

Summary:

- Logistic regression model has no skill or less skill than the climatology
- When using all months, almost no difference between using different sets of variables
- When using Oct-Mar, using only soil moisture variables shows a really small improvement

Neural Network Classification



NN Classification BSS:

All Vars, All Months, All MJO Phases: **0.15**

- Use of GPCP spatial average as target results in somewhat improved predictions
- Layers: 1 (deep network gives very large RMSE)
- Neurons per layer: 16
- Learning rate: Variable (Inverse Time Decay)
- Epochs: 500
- Optimizer: Adam
- Loss function: Sparse Categorical Cross Entropy
- L2 regularization applied