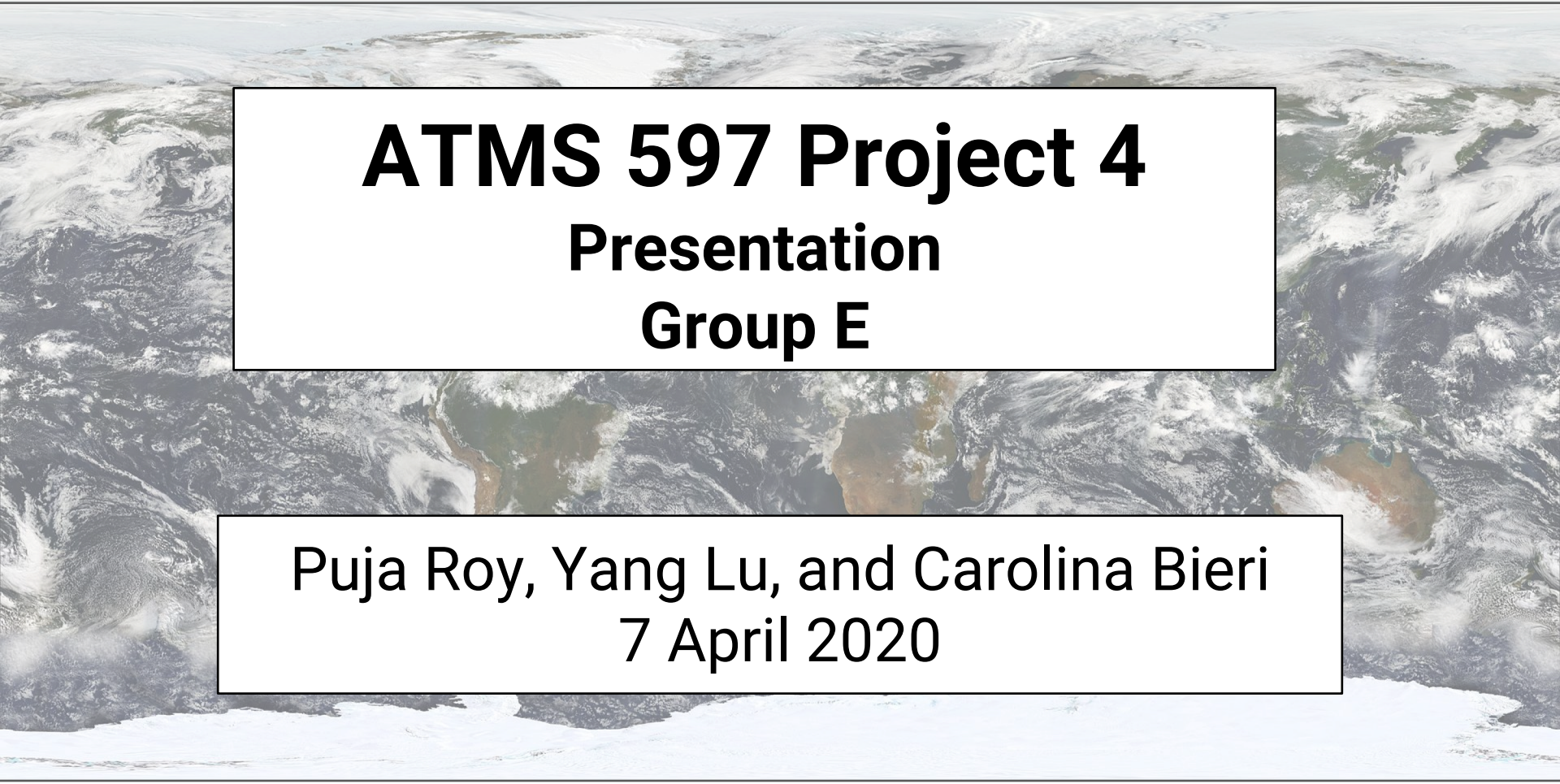


ATMS 597 Project 4
Presentation
Group E

Puja Roy, Yang Lu, and Carolina Bieri
7 April 2020

Data used

Validation: KCMI daily data (2019)

Available training data:

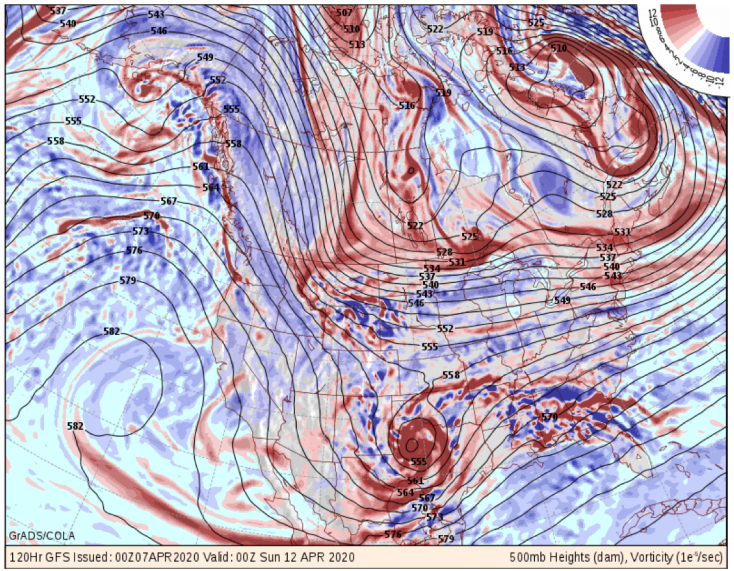
KCMI daily data (2010-2018)

GFS daily data (2010-2018)
- TMAX, TMIN, WMAX, RTOT

GFS 3-hourly data (2010-2018)
- Dewpoint (DWPC)
- Temperature (TMPC)
- Wind speed (WSPD)
- Wind u-component (UWND)
- Wind v-component (VWND)
- Geopotential height (HGHT)
- Surface pressure (PRES)
- Cloud cover (HCLD, MCLD, LCLD)

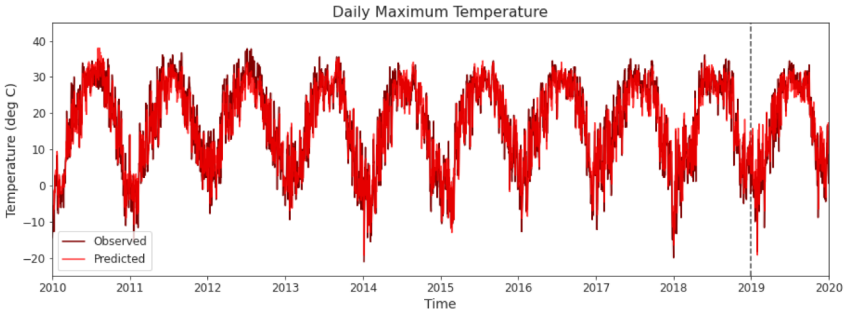
Data processing & methods

- Hourly precipitation data aggregated to daily and used for training/validation
- Missing values filled using pandas functions bfill() and ffill()
- 3-hourly GFS data aggregated to daily values
- Multiple approaches in determining predictors
  - What is simple?
  - What information is available?
  - Which features have the most weight?

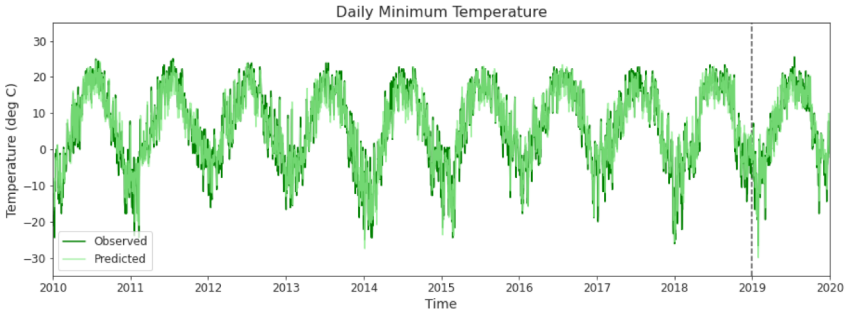


How little information can we use?

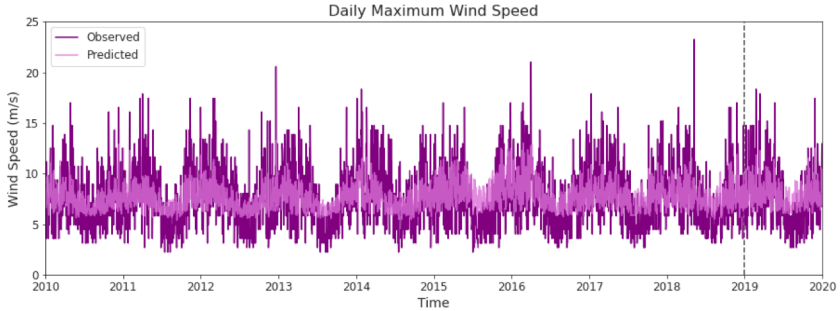
Results from “simple” model - linear regression



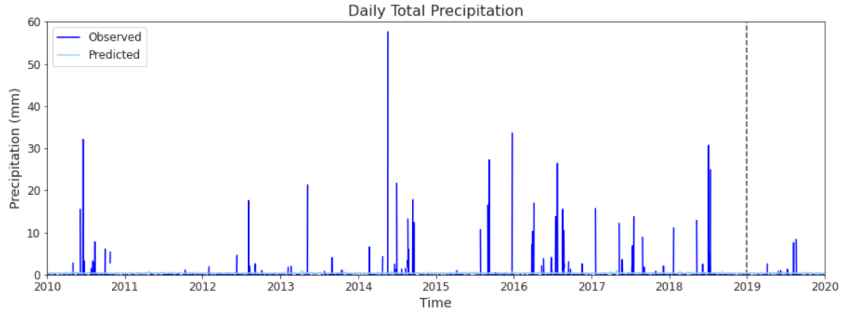
RMSE training: 4.68 °C; **RMSE 2019:** 4.86 °C



RMSE training: 3.85 °C; **RMSE 2019:** 4.01 °C



RMSE training: 2.65 m/s; **RMSE 2019:** 2.75 m/s

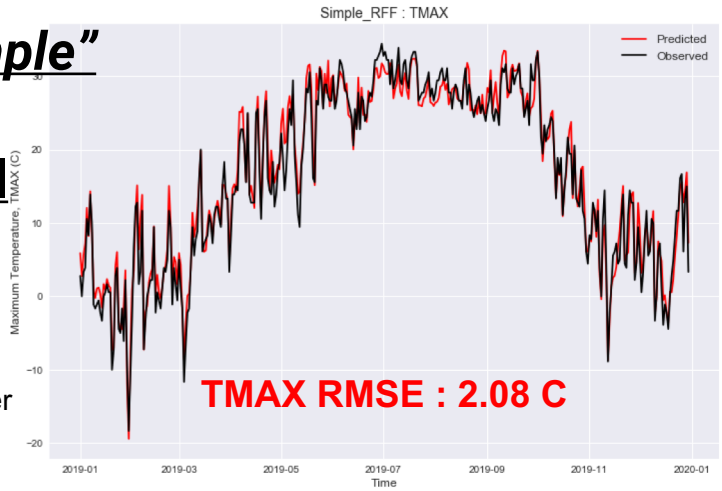


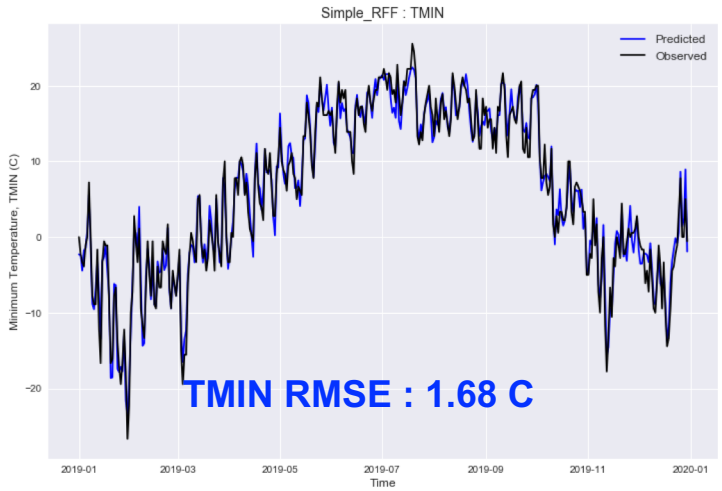
RMSE training: 2.66 mm; **RMSE 2019:** 1.64 mm

Results from *"simple"* Random Forest Regression model

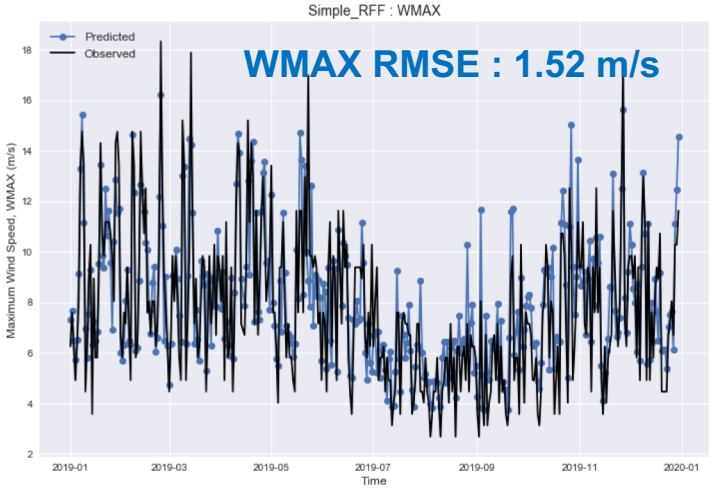
Optimized Hyperparameter settings (Used RandomizedSearchCV method, defining a grid of hyperparameter ranges, and randomly sampling from the grid and performing K-Fold CV with different combinations, the best results were achieved using the following parameters**)**

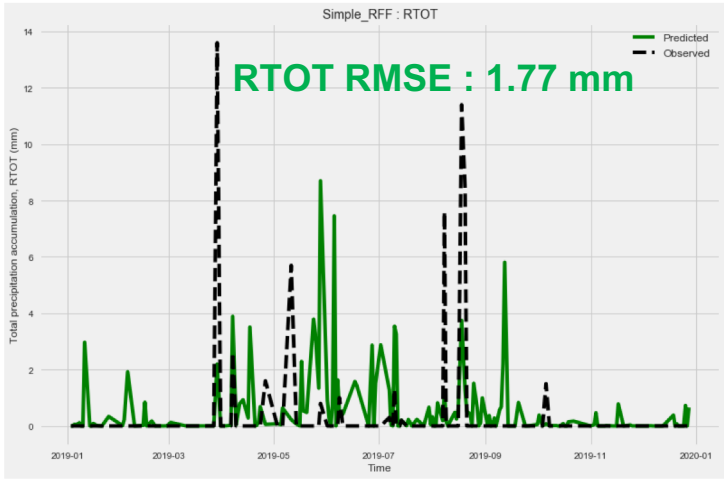
- n_estimators = 1000
- random_state = 42
- criterion = 'mse'
- max_depth = None
- min_samples_split = 2
- min_samples_leaf = 1
- min_weight_fraction_leaf = 0.0
- max_features = 'auto'
- max_leaf_nodes = None
- bootstrap = True



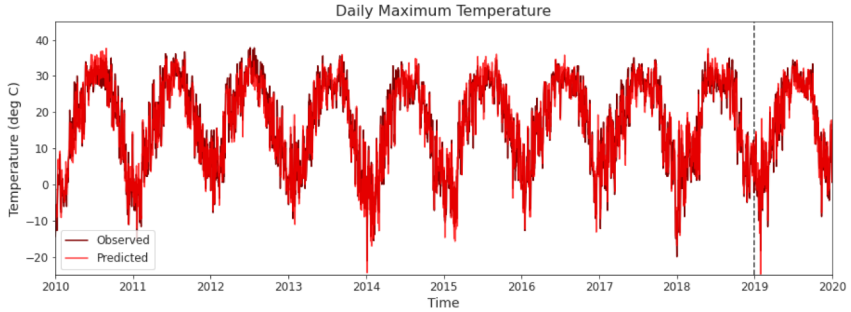


Root Mean Squared Error = RMSE

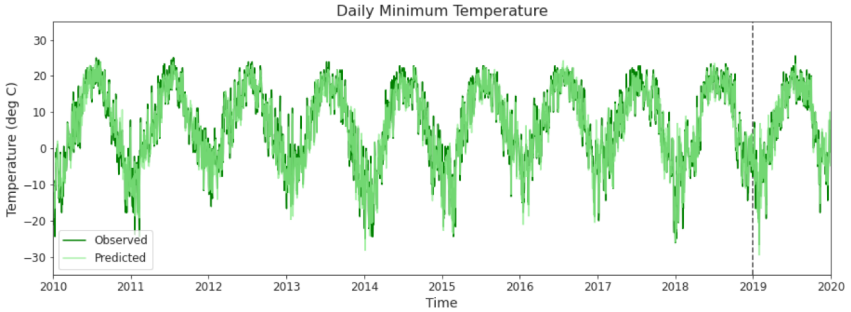




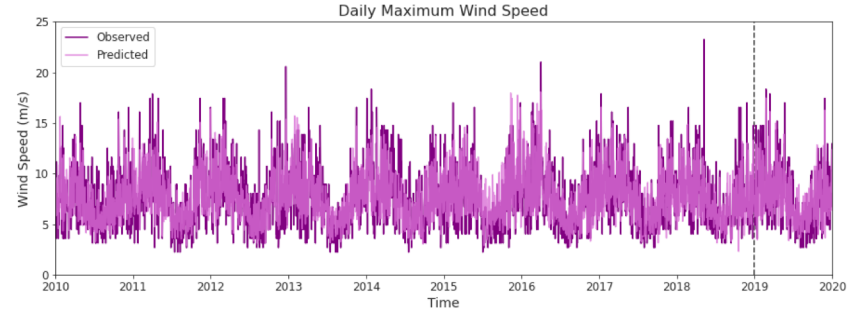
Does adding more information necessarily improve the prediction?



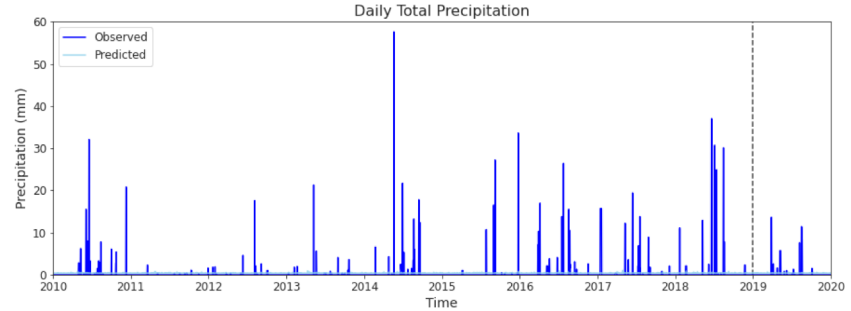
RMSE training: 2.31 °C; **RMSE 2019:** 2.41 °C



RMSE training: 2.17 °C; **RMSE 2019:** 2.40 °C



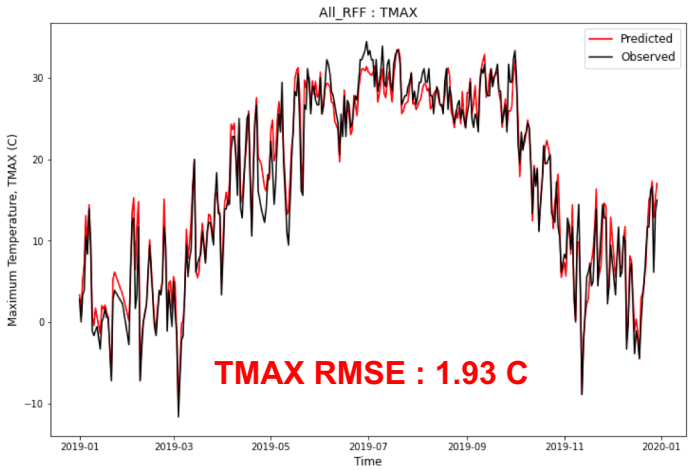
RMSE training: 1.74 m/s; **RMSE 2019:** 1.79 m/s

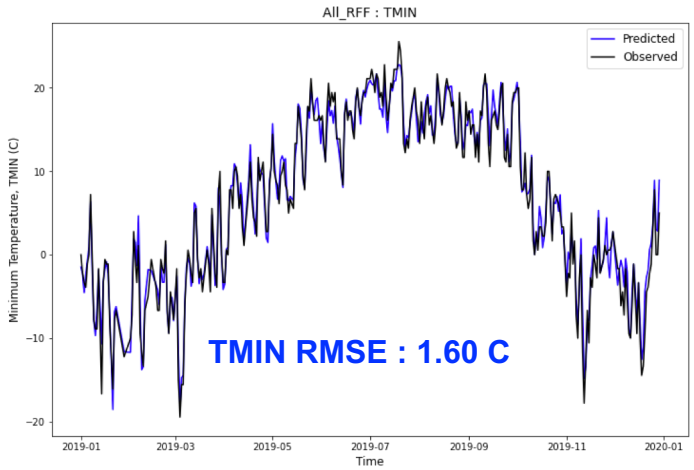


RMSE training: 2.61 mm; **RMSE 2019:** 1.62 mm

Does adding more information necessarily improve the prediction?

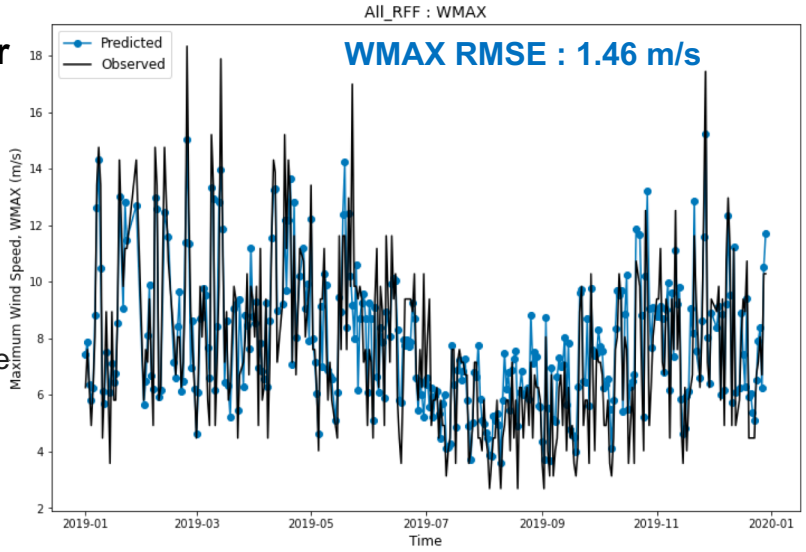
Results from *"All"* Random Forest Regression model (Missing values removed)

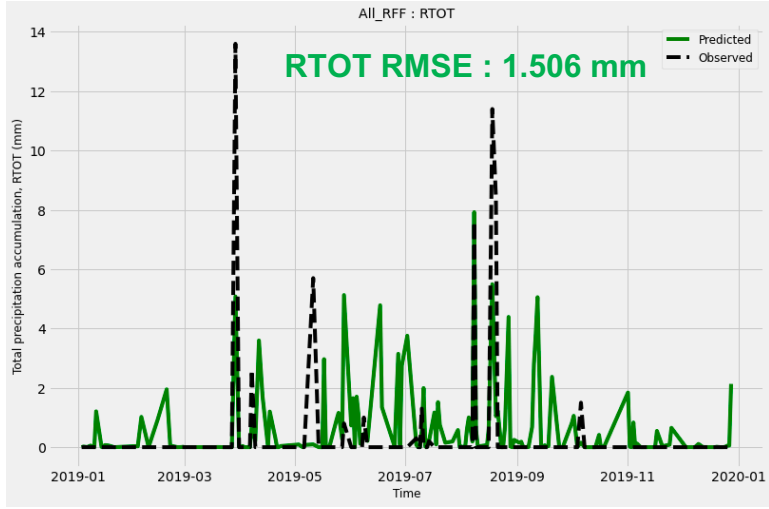




Optimized Hyperparameter settings

- n_estimators = 1000
- random_state = 42
- criterion = 'mse'
- max_depth = None
- min_samples_split = 2
- min_samples_leaf = 1
- min_weight_fraction_leaf = 0.0
- max_features = 'auto'
- max_leaf_nodes = None
- bootstrap = True

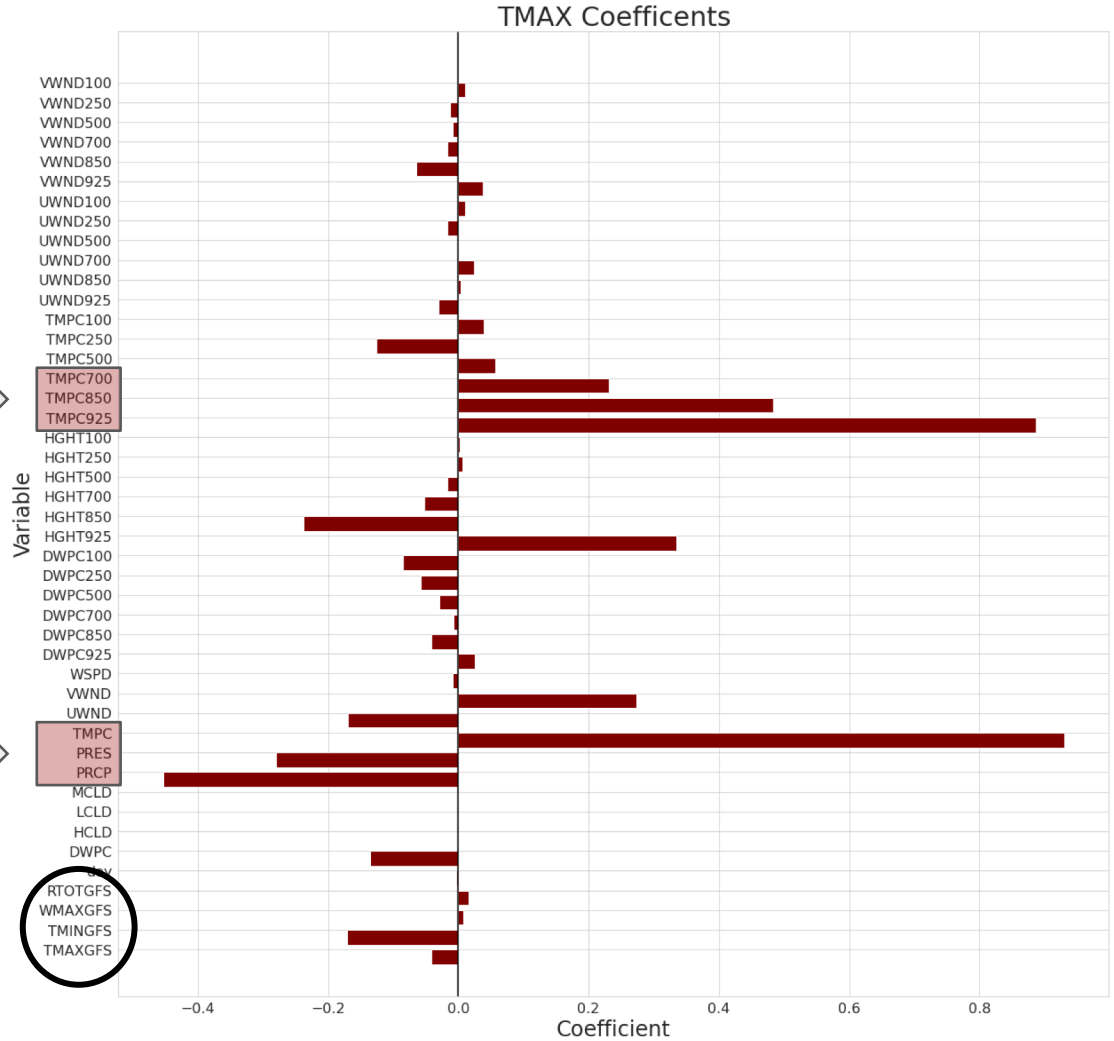




Which variables are the best predictors?

Influence of temperatures aloft

Influence of midlatitude cyclones?

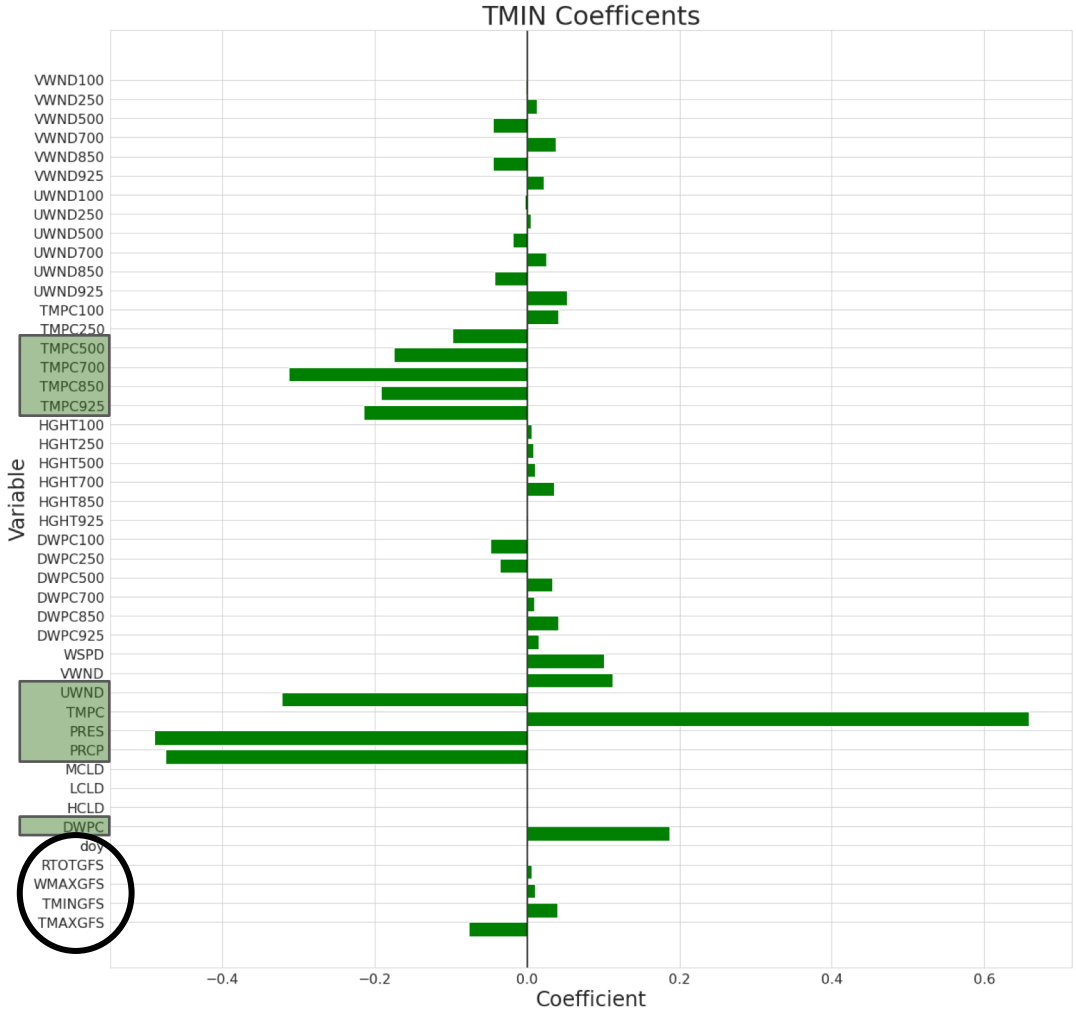


Which variables are the best predictors?

Influence of temperatures aloft

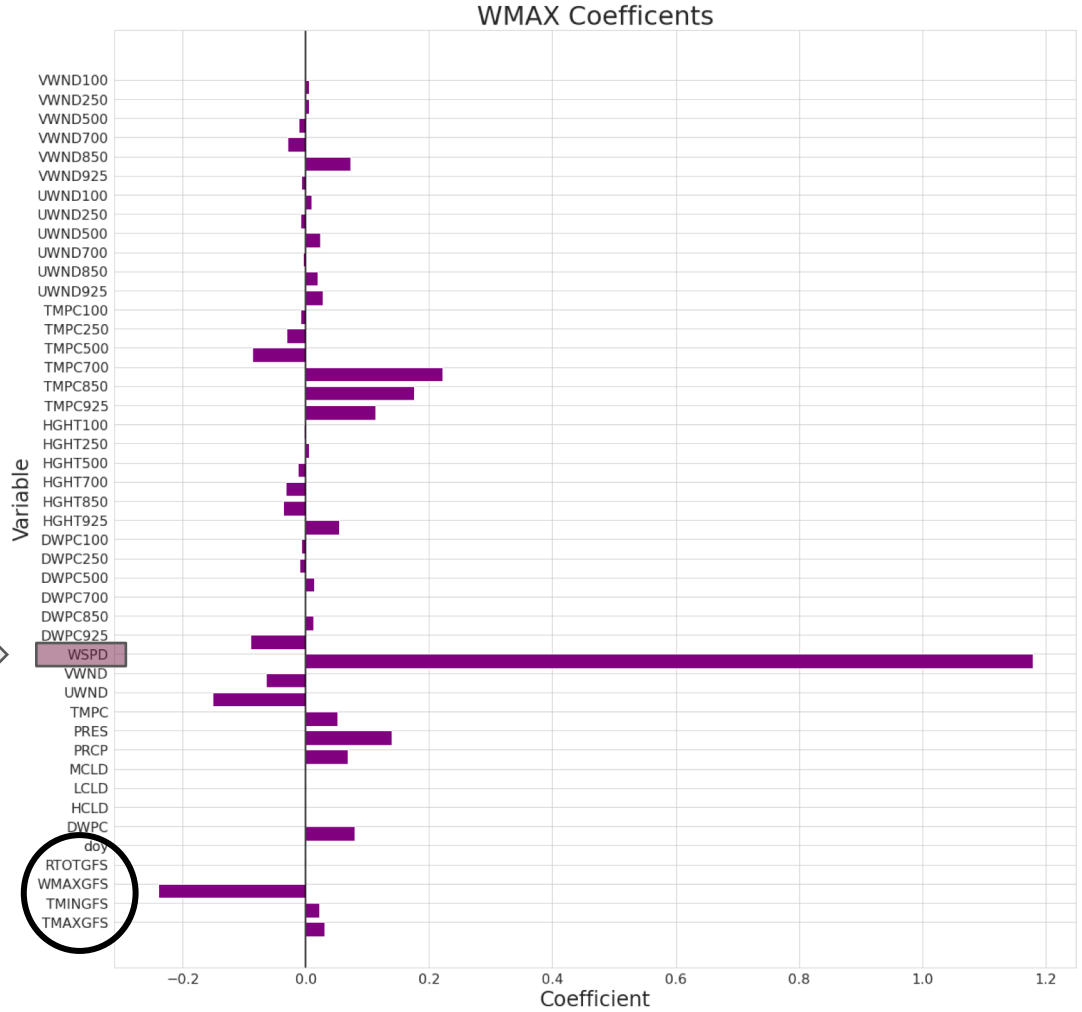
Influence of midlatitude cyclones?

Added influence of dewpoint



Which variables are the best predictors?

Influence of daily mean wind speed

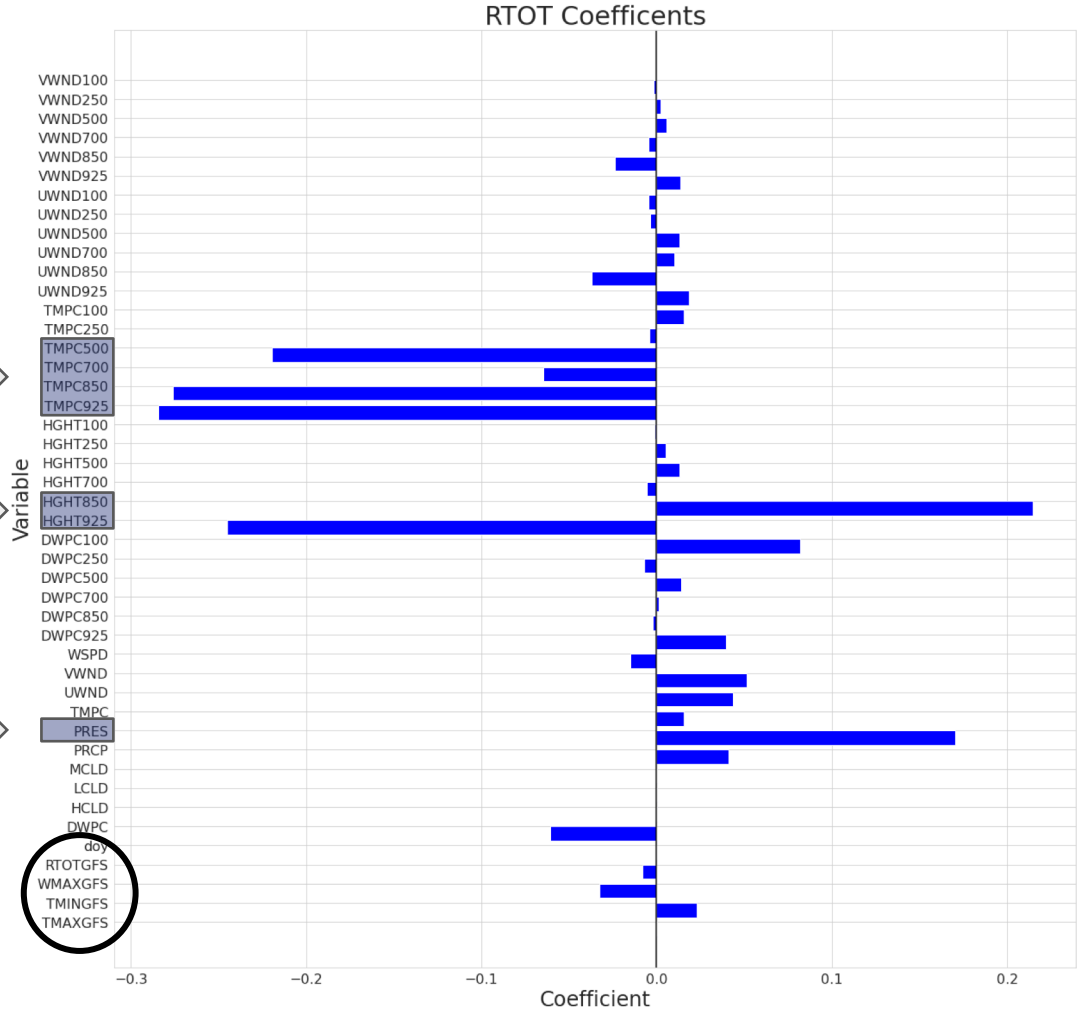


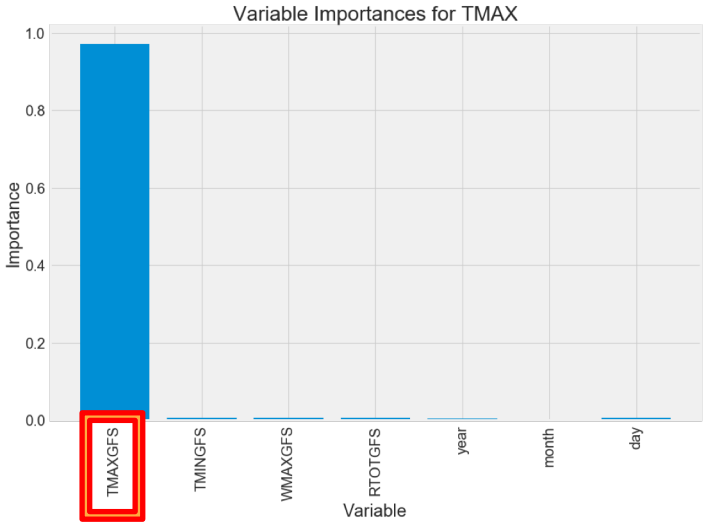
Which variables are the best predictors?

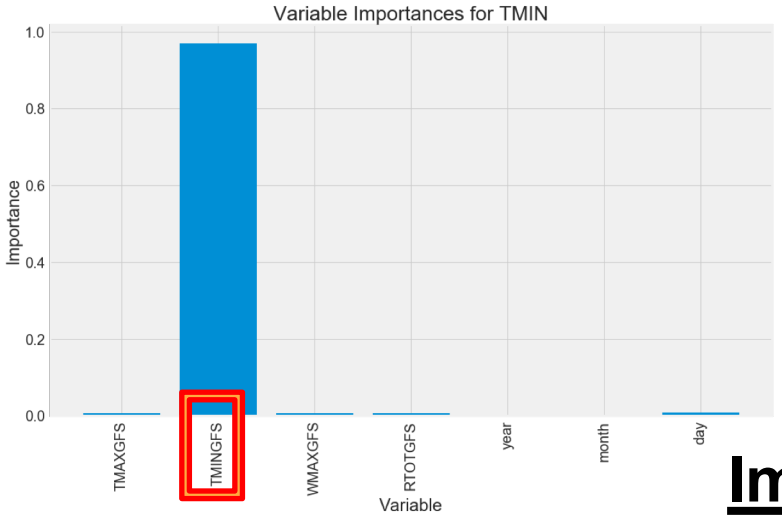
Influence of temperatures aloft - related to HGHT?

Influence of synoptic patterns

Convergence/lift

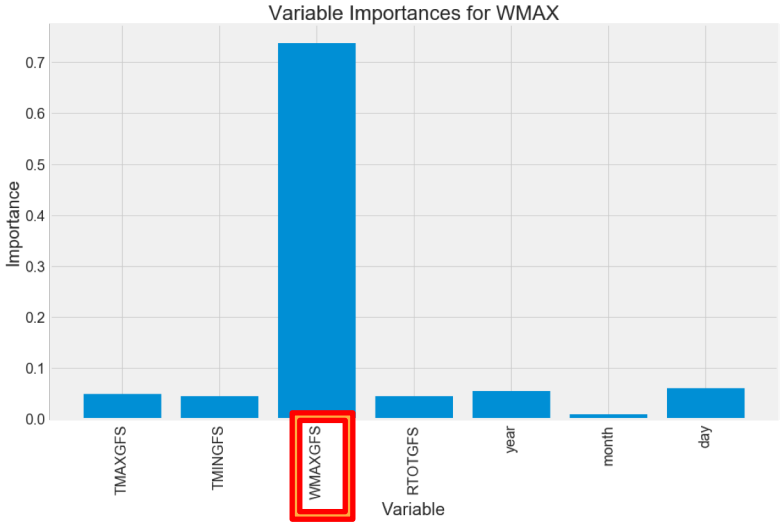


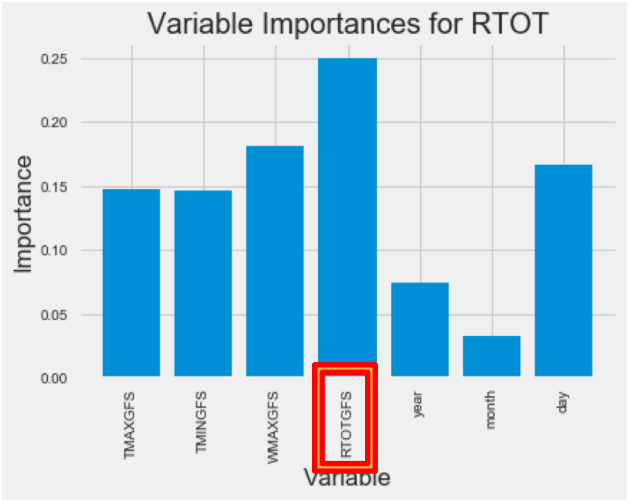


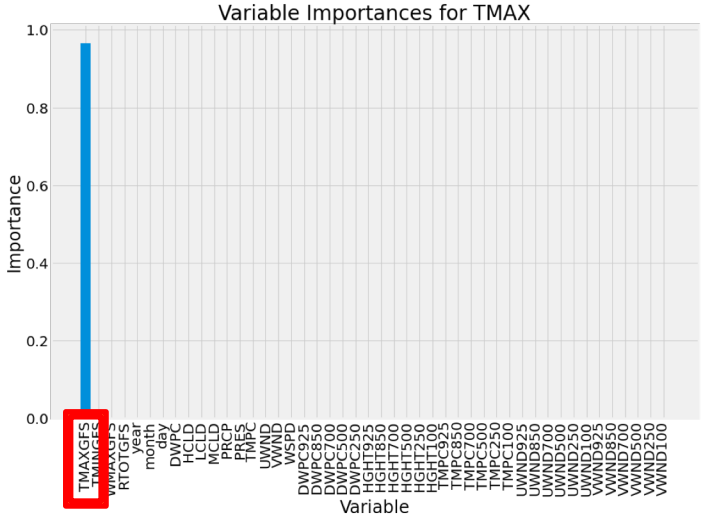


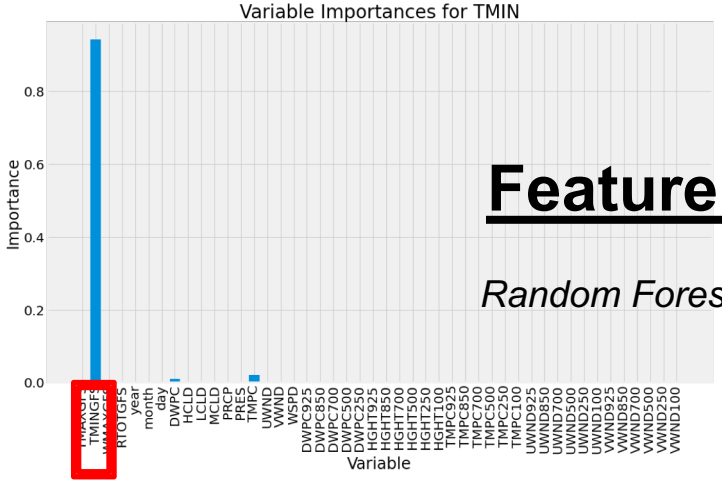
Feature Importances

Random Forest Regression "Simple" Model



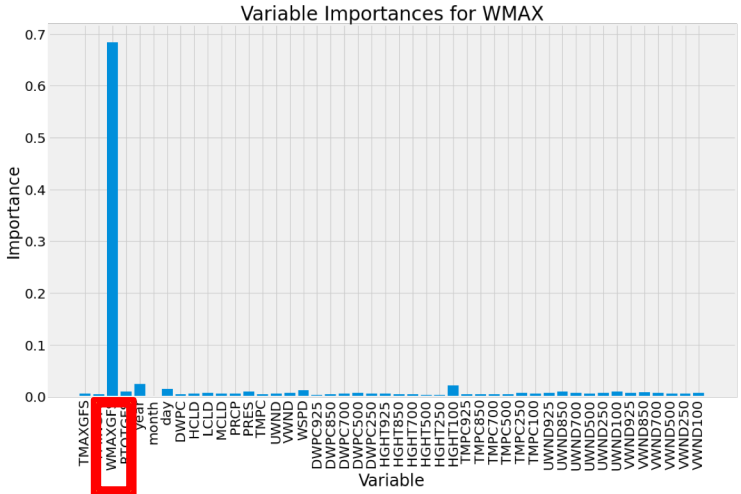


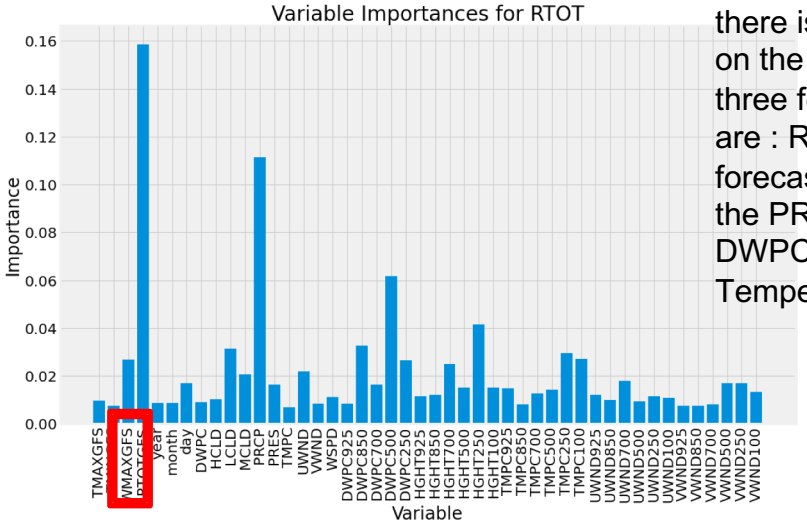




Feature Importances

Random Forest Regression "All" Model





For Precipitation, overall there is weak dependence on the features but still, the three features that stand out are : RTOTGFS (GFS forecast for precip), precip at the PRCP(surface), DWPC500 (Dewpoint Temperature at 500 hPa)

I. *Tuning the parameters to improve the model*

No.of estimators /RMSE variables	N = 10	N=100	N=1000
TMAX (C)	2.13	2.08	2.16
TMIN (C)	1.86	1.71	1.69
WMAX (m/s)	1.54	1.53	1.61
RTOT (mm)	2.35	1.73	1.76

The runs are for the "Simple" Model

II. Tuning the parameters to improve the model

RMSE after choosing the "most important" feature for the "Simple" Models:

Model Type/RMSE variables	Simple_Model, N=1000	Simple_Most_Important_Model, N = 1000
TMAX (C)	2.08	2.27
TMIN (C)	1.68	1.82
WMAX (m/s)	1.52	1.98
RTOT (mm)	1.77	1.96

III. Adding more "predictors" to improve the model

Adding all the available features(including profile and surface variables) for training the model

Model Type/RMSE variables	Simple_Model, N=1000	All_Model, N = 1000
TMAX (C)	2.08	1.93
TMIN (C)	1.68	1.60
WMAX (m/s)	1.52	1.46
RTOT (mm)	1.77	1.50

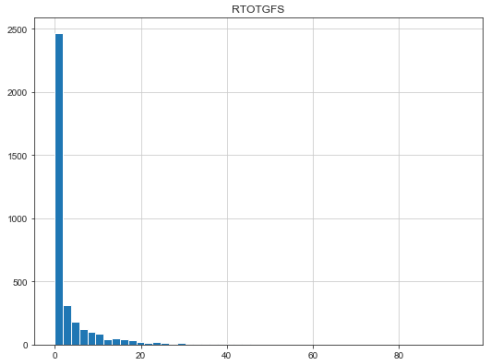
Challenges & possible improvements

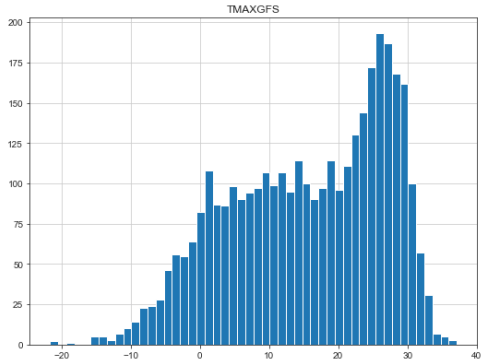
- Figuring out the best way to use available information
- Use of persistence
- More systematic way of determining predictors
- For the Random Forest Regression (RFR) Model for this case, the predictions improved after increasing the number of decision trees (n_estimators) from n=10 to 100, but not from 100 to 1000 for the “simple” runs in the model.
- The most important predictors were usually the respective GFS forecast variables - TMAXGFS, TMINGFS, WMAXGFS and RTOTGFS.
- Both for LR and RFR model - adding more predictors seemed to yield the best result
- Choosing the “most important” feature(s) didn’t improve the quality of the predictions.

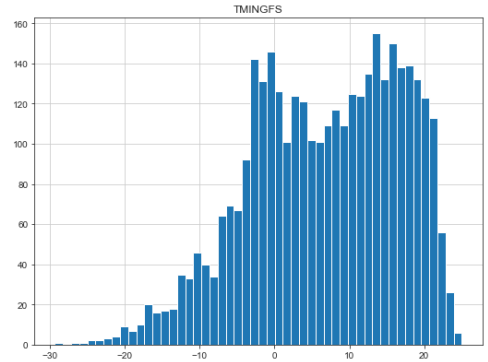
Regularization to improve the Multiple Linear Regression Model

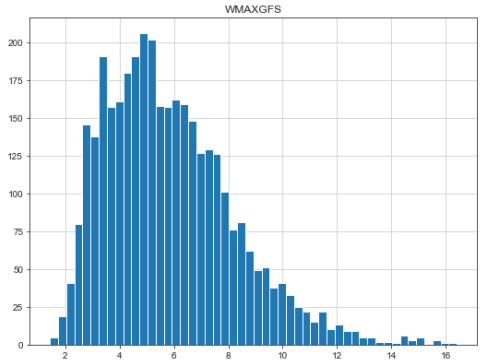
Data preprocessing

Dropped the “DWPC100” column, dropped all other NaNs



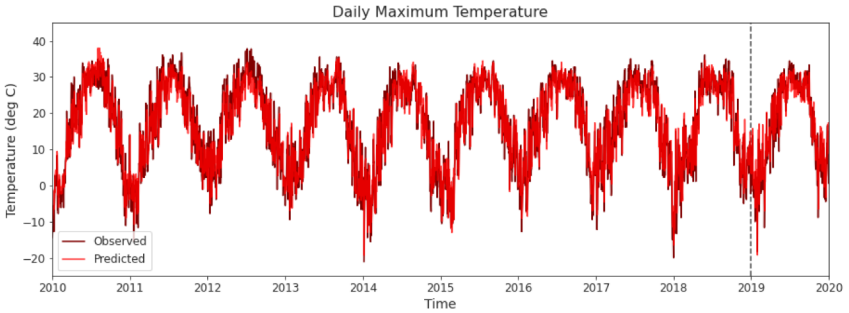




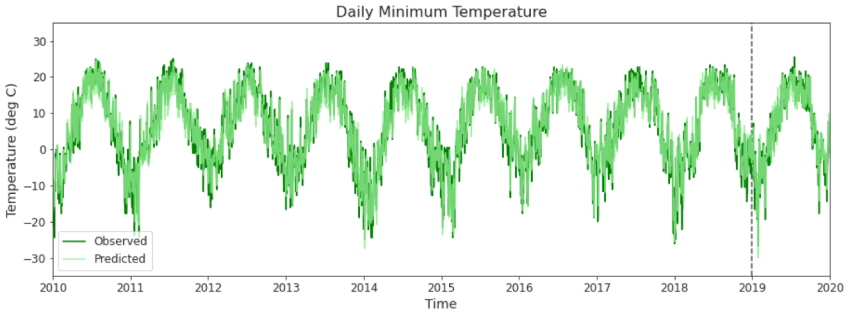


How little information can we use?

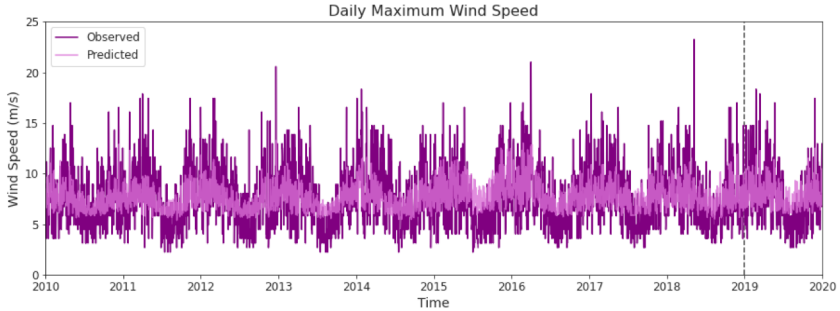
Results from “simple” model - linear regression



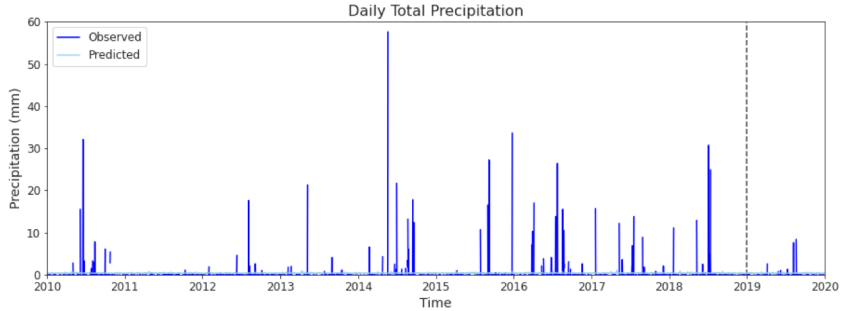
RMSE training: 4.68 °C; **RMSE 2019:** 4.86 °C



RMSE training: 3.85 °C; **RMSE 2019:** 4.01 °C



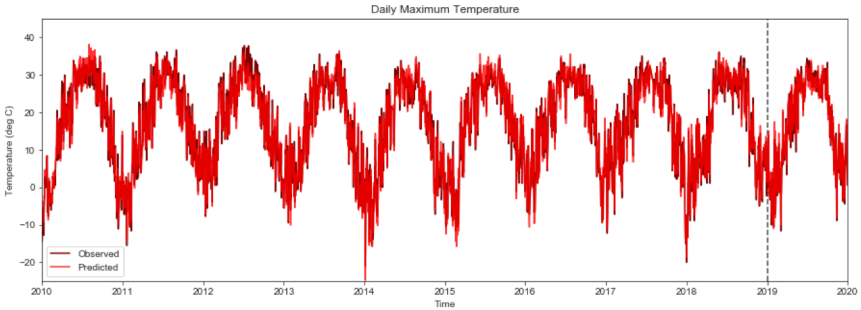
RMSE training: 2.65 m/s; **RMSE 2019:** 2.75 m/s



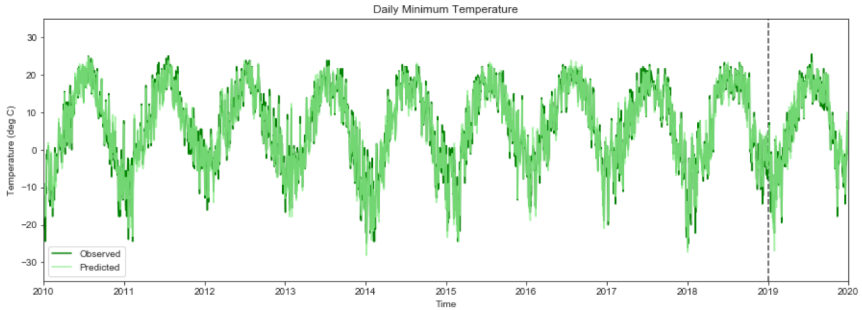
RMSE training: 2.66 mm; **RMSE 2019:** 1.64 mm

Linear regression

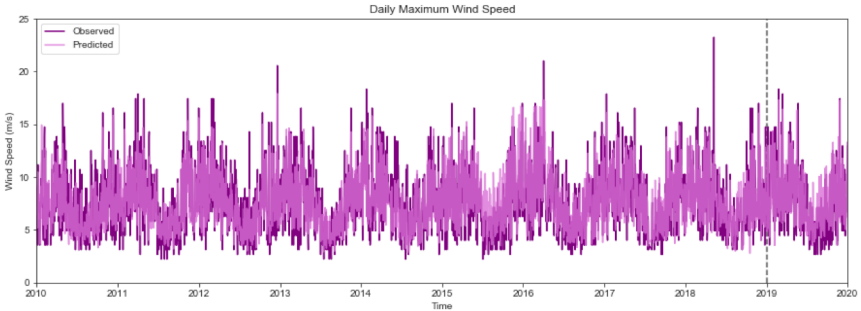
With averaged 3hr data



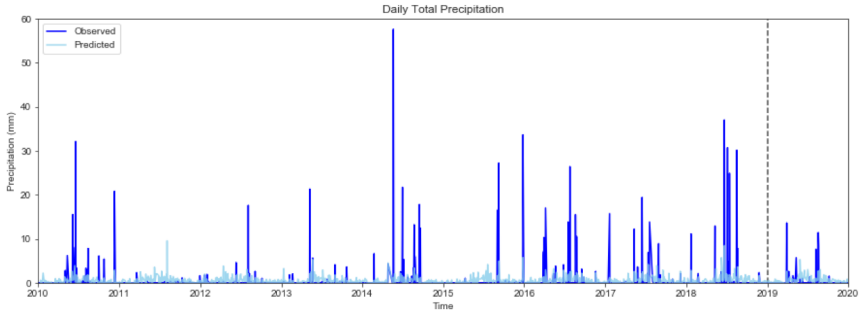
RMSE training: 1.93 °C; **RMSE 2019:** 1.91 °C



RMSE training: 1.68 °C; **RMSE 2019:** 1.73 °C



RMSE training: 1.50 m/s; **RMSE 2019:** 1.58 m/s



RMSE training: 2.85 mm; **RMSE 2019:** 1.52 mm

Regularization on Linear regression

With averaged 3hr data– Lasso, Ridge and Elastic net

About 3/7 of total precipitation observations are missing. Regularization improves the robustness of regression. Lasso is too strong, but all regularizations improve the prediction of total precipitation slightly.

Regularization	TMAX19	TMIN19	WMAX	RTOT
None	1.91	1.73	1.58	1.65
Lasso	11.71	10.65	2.98	1.59
Ridge	2.92	2.62	1.76	1.52
Elastic Net (R:L=1:9)	11.66	10.62	2.98	1.59

Using more raw data helps little

- Since we average 3-hourly data into daily means, lots of information are lost.
- To preserve the daily variance of the predictors, we regroup the 3-hourly data to daytime means (9:00 to 18:00) and nighttime means (0:00 to 6:00 and 21:00)

- The results show that adding more features don’t affect the estimation from linear regression much. For random forest regression, the estimation of TMAX gains a marginal improvement, RMSE drops from 1.93 °C to 1.91 °C.

- Further separation of features won’t do much good for this task.

Hyperparameters for regression forest

- We searched for the best hyperparameter using GridSearchCV. 3-fold cross-validation is used during the following test
- The effect of increasing number of estimators enters a plateau after 1000.
- The minimum samples to split is 8, however the improvement is negligible compared with the default, 2.
- For this task, we leave other hyperparameters to be unlimited.

Thank you!

Questions?