Estimation of Daily Water Temperature using Random Forest

B Steele

October 5, 2023

GH Repo

Scientific motivation and problem statement:

Water temperature is often an indicator of water quality, as it governs much of the biological activity in freshwater. While temperature is an important parameter to monitor in freshwater lakes, manual monitoring of waterbodies (by physically visiting a site) and sensor networks to monitor water temperature, are costly endeavors.

In this example, I will use Random Forest to estimate water surface temperature for reservoirs with long manual monitoring data from Northern Water. The features that I will be using to estimate surface temperature include summary NLDAS meteorological data (air temperature, precipitation, solar radiation, and wind) as well as static values for each of the reservoirs (elevation, surface area, maximum depth, volume, and shoreline distance).

The comparative baseline for this analysis will be the day-of-year average water temperature across all lakes and years. The baseline estimates result in a MAE of 2.24 deg C and MSE of 2.74 deg C.

In addition to the manual sampling record that is maintained by Northern Water (n = 1125), I will be leveraging surface temperature estimates from the Landsat constellation, Landsat 4-9 (n = 5039). These thermal estimates are well-aligned with the manual monitoring data for the 7 reservoirs and have been bias-corrected for over estimates in the warmest months. 'Surface temperature' in the manual sampling record for this example is any measured temperature at >= 1m depth. I retain only the top-most value for temperature. Static variables are only available for 6 of 7 reservoirs, so Windy Gap reservoir has been dropped from this analysis.

Table 1: Static variables used in the Random Forest algorithm. Windy Gap Reservoir has incomplete data and has been dropped from this analysis.

feature	elevation	area	shoreline_length	max_depth	volume
Carter Lake	5760	1100	12.0	180	112230
Grand Lake	8370	507	4.5	389	68621
Granby Reservoir	8280	7256	40.0	221	530000
Horsetooth Reservoir	5430	1850	25.0	200	150000
Shadow Mountain Reservoir	8367	1346	8.0	24	16800
Willow Creek Reservoir	5400	303	7.0	124	10500
Windy Gap Reservoir	NA	NA	NA	NA	445

Eventual implementation of this algorithm will include forecasting of temperature for these lakes as well as lakes that have only Landsat-derived temperature estimates and that are not included in this dataset. Because I want this algorithm to perform well on new lakes, I want to take steps to make sure that the algorithm is not overfit to these specific lakes.

No pre-processing (i.e. regularization) was completed for these data, as decision trees make purely empirical decisions, and that type of pre-processing is not usually necessary. I have pre-processed the NLDAS data to provide summaries of the previous day weather, 3 days prior, and 5 days prior - meaning, the model does not use *today's* weather for prediction.

Training/Validation/Testing

During data exploration, it was clear that there are site-level differences in temperature range and general seasonal response for each water body. These differences are likely due to static variables that differentiate these water bodies. That said, if I add in site-level information, the algorithm may have a propensity to "learn" those key attributes and likely overfit to the data, not allowing for generalization beyond these lakes. I will need to look at the RF trees, feature importances, and permutation importances to make sure these features do drive the results of the model (which might indicate that the model is overfit to the identifying characteristics).

For training and validation I use a leave-one-out method that will result in six random forest models where each iteration will use data from a single lake for validation and the other five for training. Since the intended implementation will be daily forecasts, testing performance will be assessed through hindcasting. The hindcast dataset is a holdout dataset beginning in 2020 across all lakes.

Results

Hyper-parameter tuning

I manually iterated on hyper-parameter settings during training of the model, tyring various number of estimators (30-70), maximum tree depth (3-7), node split minimum (5-20), and leaf split minimum (5-10;). Most attempts at hyper-parameter tuning did not result in significant changes in validation performance. I generally chose more conservative hyperparameters in as a way to assure that I do not overfit the training data (estmiators = 40, maximum tree depth = 5, node split minimum = 15, leaf split = 10) resulting in MAE of the validation ranging between 1.46 and 1.79 and MSE ranging between 1.86 and 2.42.

Hindcasting application

Because all random forest models in training and validation performed similarly across the ensemble and the feature importance and permutation importance (see Supporting Figures) were similar, I've decided to aggregate all data together and create a single RF algorithm to use for hindcasting. In an effort to make sure that this RF is built similarly to the previous leave-one-out models, I'll examine the RF tree, feature importance, and permutation importance.

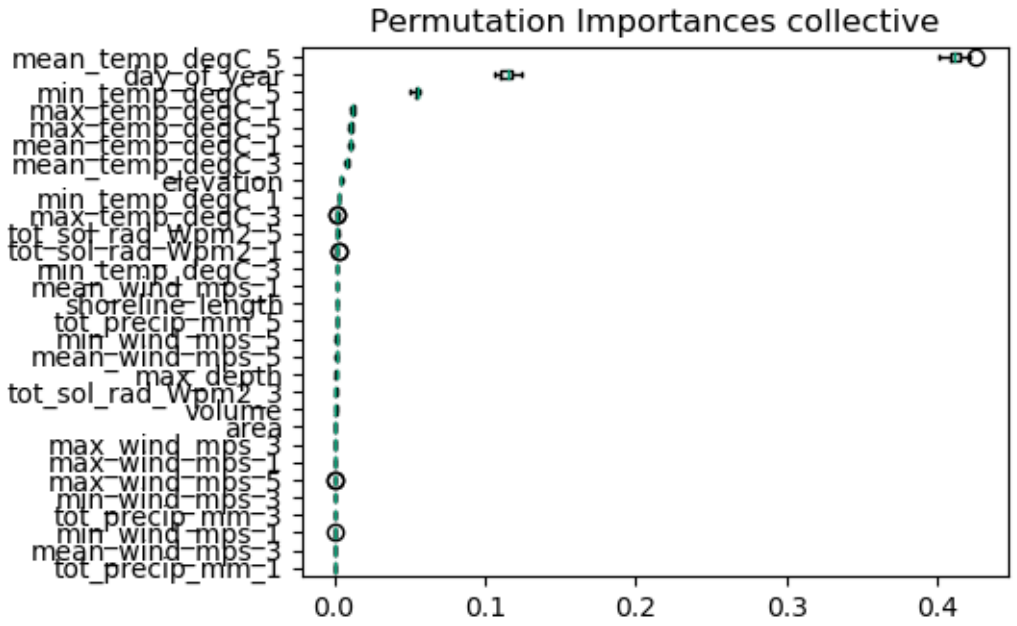

Figure 1: Permutation importance for variables in the hold out testing dataset.

This iteration of the model shares similar feature importances to the leave-one-out models, indicating that this model architecture is similar to the previous and does not seem to fit according to the static variables. Whether or not this algorithm is overfit to the training data is unclear. The hindcast test dataset had a higer MAE (1.99) and MSE (2.45) than the leave-one-out validation datasets. Additionally, the collective training MAE (1.31) and MSE (1.75) was considerably lower than the test MAE and MSE. This indicates to me that the collective training decision may have resulted in overfitting of the data.

Discussion

If the collective model is not overfit, it is possible that error is propagated from the the remote sensing data, climatological patterns, or inadequate training features. The remote sensing data has about a 1 deg C error associated with it. 2020 forward was a particularly dry time in the climatological past - most of Colorado experience severe drought and many reservoirs were at historically low levels until this past year. Finally, it is possible that the included meteorological summary features are not adequate for robust training of the algorith - inclusion of heating degree days and cooling degree days could be a useful way to embed seasonal change as well as heat/cooling persistence. This could possibly help with the early season misalignment see in Figure 2.

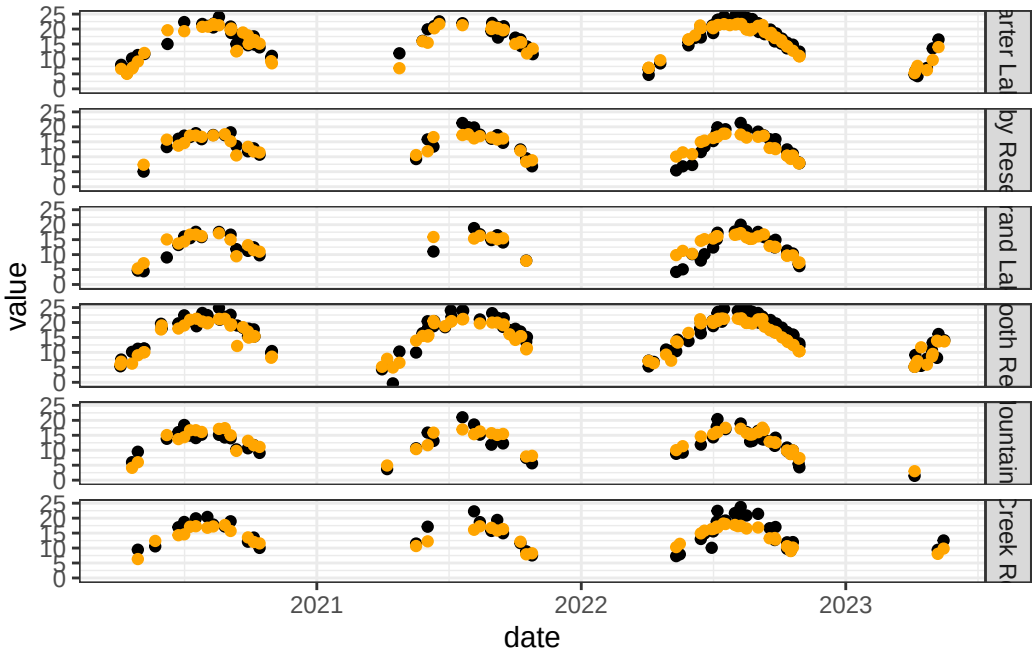

Figure 2: Datetime graph with actual values (black) and predicted values (orange), which shows the model is capturing the dirunal cycle.

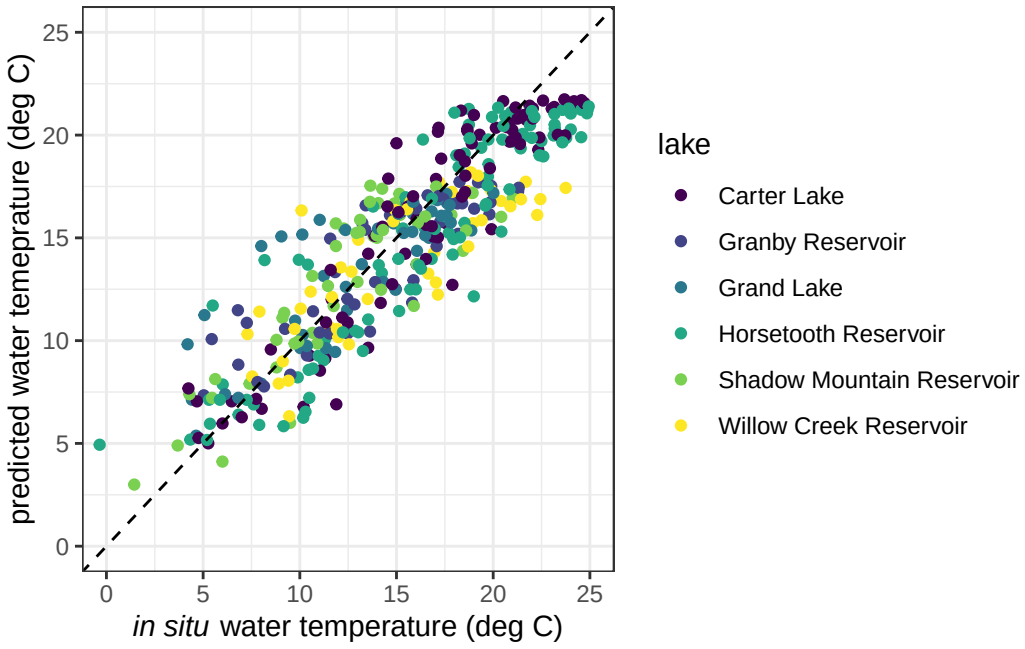

Figure 3: Scatter plot of predicted temperature and observed temperature at each of the 6 lakes in the dataset for the testing set. Black dashed line is 1:1.

Supporting Figures

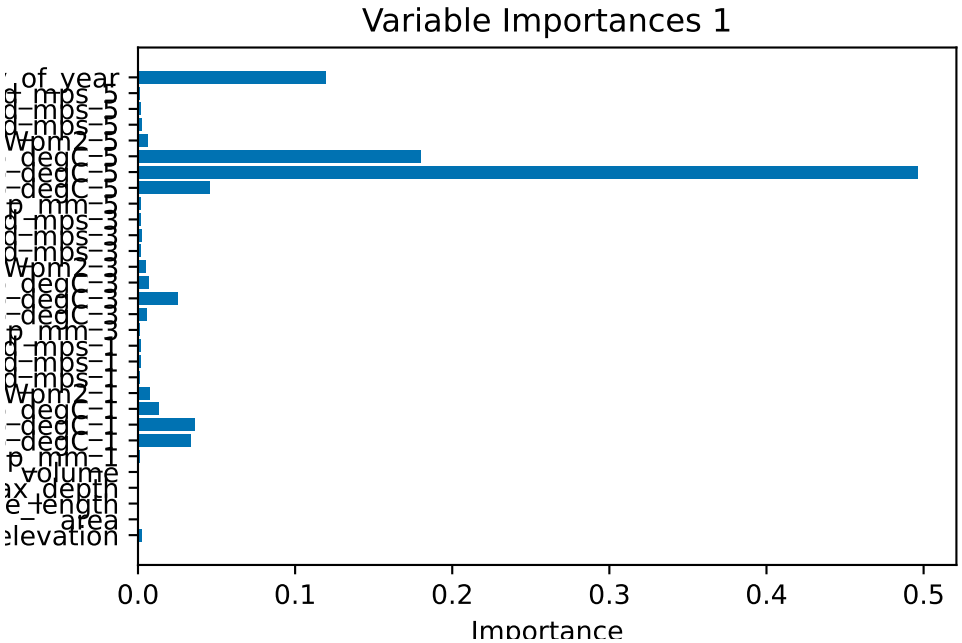

Figure 4: Feature imporances across all leave-one-out validation datasets with consistent feature imporatnces across 5-day previous summaries and minimal importance of the waterbody.

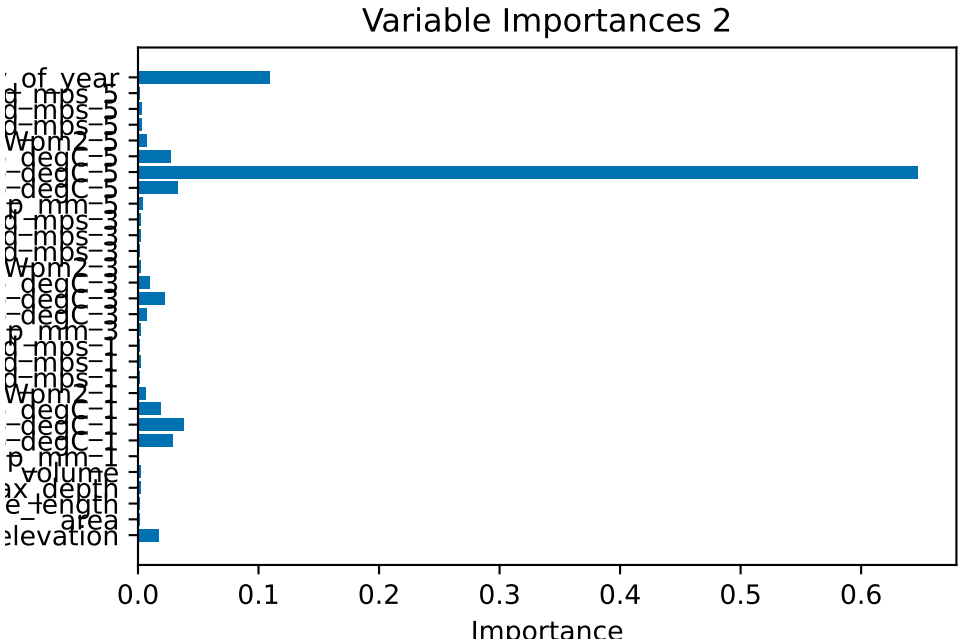

Figure 5: Feature imporances across all leave-one-out validation datasets with consistent feature imporatnces across 5-day previous summaries and minimal importance of the waterbody.

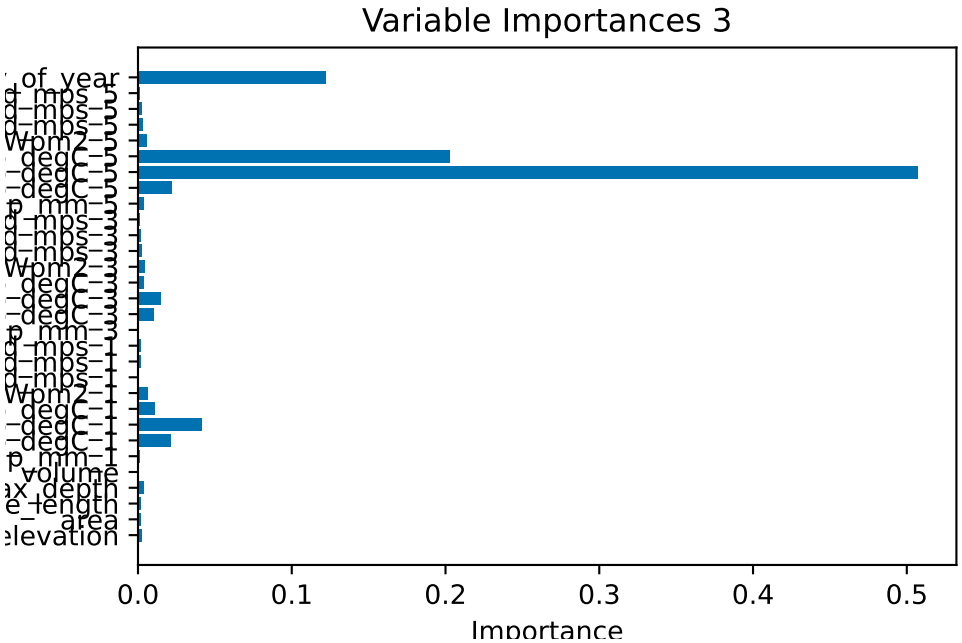

Figure 6: Feature imporances across all leave-one-out validation datasets with consistent feature imporatnces across 5-day previous summaries and minimal importance of the waterbody.

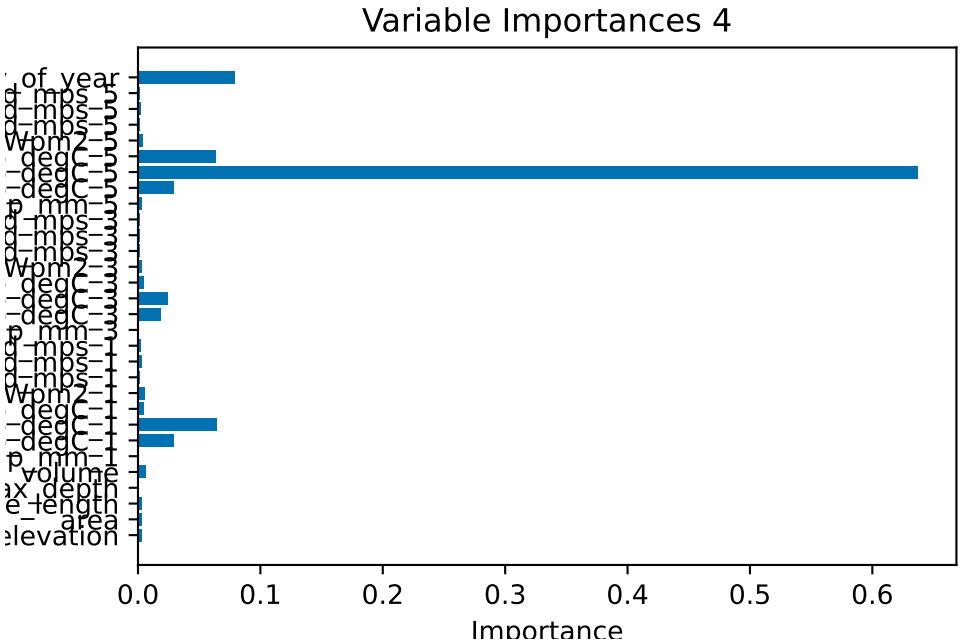

Figure 7: Feature imporances across all leave-one-out validation datasets with consistent feature imporatnces across 5-day previous summaries and minimal importance of the waterbody.

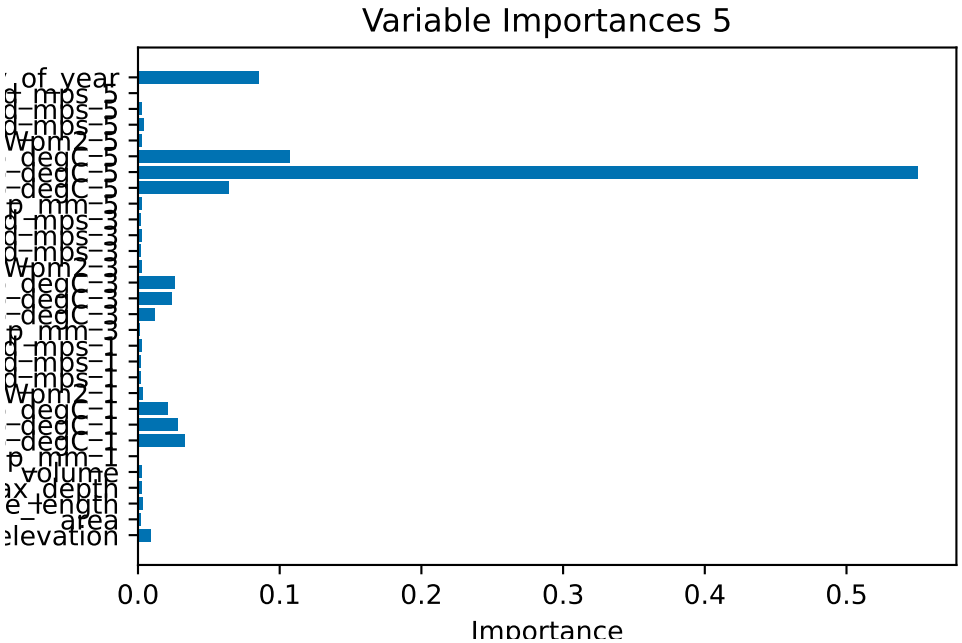

Figure 8: Feature imporances across all leave-one-out validation datasets with consistent feature imporatnces across 5-day previous summaries and minimal importance of the waterbody.

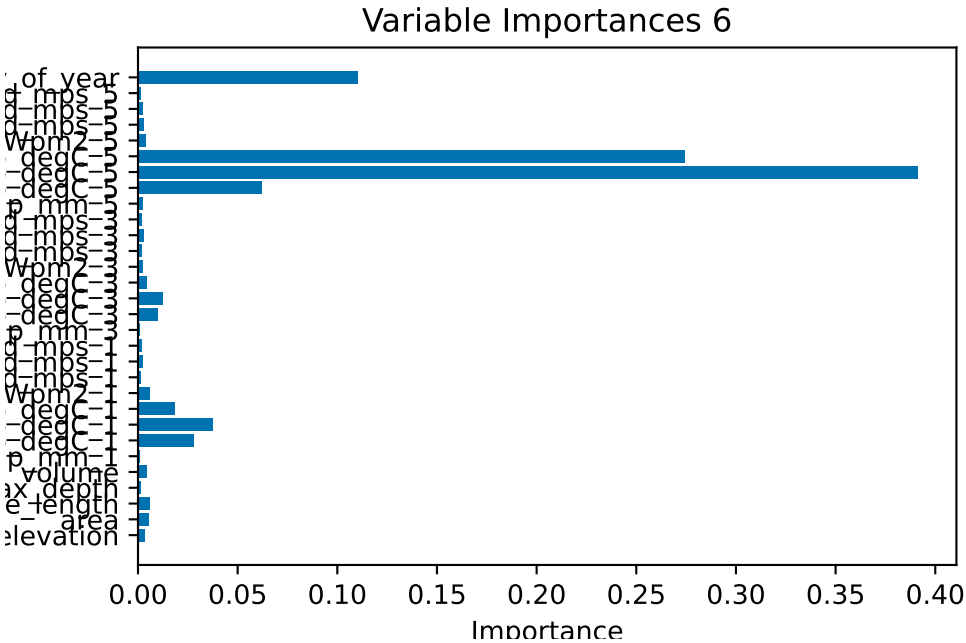

Figure 9: Feature imporances across all leave-one-out validation datasets with consistent feature imporatnces across 5-day previous summaries and minimal importance of the waterbody.