Portland State UNIVERSITY

USGS Water Quality

Jared Goldsmith, Leif Watkins, Mina Mehdinia, Savita Upadhyay

May 8, 2023

Contents

0.1 Executive Summary

This study examines data from the Klamath River at Keno and Miller to determine the impact of weather conditions and river flow on the thermal stratification of the water column. We recommend employing generalized additive models and random forest models for this purpose. Furthermore, either k-means clustering or the distribution of the temperature difference variable can be utilized to establish a threshold for thermal stratification. Our findings, however, do not result in a full solution and analysis due to the incomplete nature of the dataset utilized.

0.2 Introduction and Background

The present study aims to analyze water quality data collected between 2016 and 2021 from two continuous water-quality monitors located along the upper Klamath River. It flows 250 miles from Keno dam to the Pacific Ocean. The upstream reach including Keno and Miller are the most polluted reach. Sometimes oxygen in this stretch can go to zero which is not good for aquatic life, and Fish kills happens.

It was in the news that Federal Energy Regulatory Commission has approved removing some dams from the downstream of the Klamath River. When this happens, they are expecting to repopulate Salmon in the upper region, and they are expecting it to head water stream. But the middle stretch is a blocker because of poor water quality.
This reach is outfitted with some continuous monitor to look at the water quality over time. There are some wastewater treatment plants near Klamath Falls as well as agricultural inputs. However, the big water quality driver is the algae from the upper Klamath Lake. Dense bloom in the summer that flows up into this river. In the middle stretch the algae decay and die and settle at the bottom and in this process, they consume a lot of oxygen. There is no algae data available for the current study.(1)

0.2.1 Some key terms:

Thermal Stratification:

Thermal stratification is the vertical layering of water in a body of water with different temperatures, which leads to a depletion of oxygen in the lower layer and creates a toxic environment, but factors like wind and large flows can reduce the toxic environment and increase dissolved oxygen, making it crucial to understand and manage thermal stratification for river health.

Auto-correlation:

Auto-correlation measures the correlation between current and past/future values in a time series, and it's important to consider in statistical analyses as it can affect the validity of results and lead to incorrect conclusions; ACF plot, PACF plot, and Durbin-Watson tests are some tools used to check auto-correlation.

0.2.2 Goal

The primary objective of this study is to develop methods and models that address several key questions related to the thermal stratification of the water column in the upper Klamath River.
The questions being posed include:

- How do various weather conditions near the Klamath River and river flow influence the thermal stratification of the water column?
- What constitutes an appropriate threshold level for defining thermal stratification?
- What are the weather conditions and approximate dates of the year when reverse thermal stratification occurs, as well as the approximate dates when the stratification reverts to positive?
- At which hours of the day is temperature stratification at its peak, and when is the river at its most unstratified state?

By establishing appropriate methods to address these questions, the research seeks to enhance our understanding of the thermal stratification dynamics within the upper Klamath River.

0.3 Data Acquisition and Processing

The United States Geological Survey (USGS) provided four distinct datasets for this study. The first dataset, sourced from the Agrimet program managed by the United States Bureau of Reclamation (USBR), contains detailed hourly weather data collected in close proximity to the USGS sensor locations. The Agrimet dataset includes measurements on air temperature, relative humidity, wind direction, peak wind gust, wind speed, solar radiation, and precipitation. It is essential to note that the Agrimet weather data is reported in local time (Mountain and Pacific time zones), and all temperature values they collect are in Fahrenheit.(2) However, the USGS reports temperatures in Celsius, therefore Agrimet temperatures were converted into Celsius for the purposes of this study. The second dataset, the USGS Link River flow data, provides site numbers and river flow rates measured at 15-minute intervals. The final two datasets pertain to water quality data from Miller Island and Keno Island. Both datasets were collected using two probes positioned at each site, with one probe situated near the surface and the other close to the bottom of the river. Hourly (and occasionally half-hourly) water quality measurements were gathered for both the upper and lower probes, capturing data for variables such as water temperature, specific conductance, dissolved oxygen, pH, date, and site number.

During the data preparation stage for analysis, two major challenges were addressed. The first involved identifying and managing missing values according to the client's specifications. This entailed excluding data points from the analysis if crucial information, such as water temperature, weather, or flow data, exhibited more than three consecutive missing values. All other missing data were treated using linear interpolation. The second challenge was the merging of data from multiple sources. To accomplish this, several functions were devised to process and manipulate data from the four datasets. These functions initially facilitated the merging of flow and weather data into a single dataset based on their timestamps. Additionally, water quality data was consolidated by combining upper and lower river data using shared timestamps, rounded within a five-minute interval to the nearest hour. Following this, the water data was integrated with the previously merged flow and weather data, culminating in a comprehensive dataset for both Keno and Miller data, encompassing all predictor and response variables. To examine seasonal, monthly, and hourly variations in the data, supplementary columns were generated representing these temporal variables using timestamps. Moreover, two additional columns were created: one to convert wind direction into 16 cardinal directions and another to apply the sine function to the wind direction. The performance of these transformed variables can be tested to determine which one is better suited for the models.

0.4 Exploratory data analysis

To gain a deeper understanding of the data and identify patterns, trends, and potential issues, an exploratory data analysis was performed, supplemented by various data visualization techniques. The initial step involved generating a heatmap to reveal correlations between predictor variables and to detect any multicollinearity. In Figure 1, a strong positive correlation between wind gust and wind speed can be observed, along with a strong negative correlation between humidity and air temperature, and between humidity and solar radiation for Miller dataset. Figure 8 in Appendix also shows the same result for Keno. To address the issue of multicollinearity, a generalized additive model and concurvity were employed, which will be discussed in a later section.

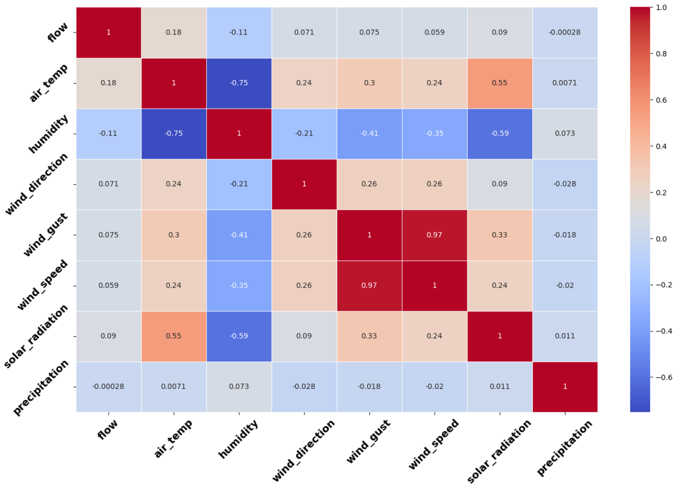


Figure 1: Correlation plot between predictors for Miller

Given the seasonal nature of the data, the dataset was divided into four distinct seasons: Winter, Spring, Summer, and Fall. Box plots were created for each season to identify any outliers that might affect the analysis. As illustrated in Figure 2 and 3, some outliers were detected for Miller dataset and also we got very similar result from Keno as illustrated in Figure 9 and 10 in Appendix; however, the client chose not to address this issue.

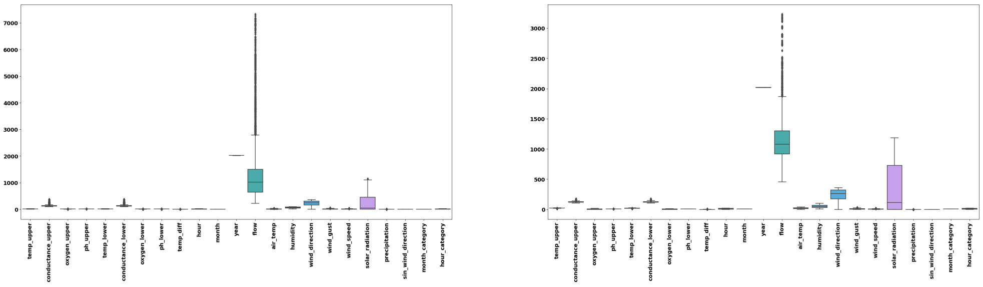


Figure 2: boxplot for Spring and Summer for Miller

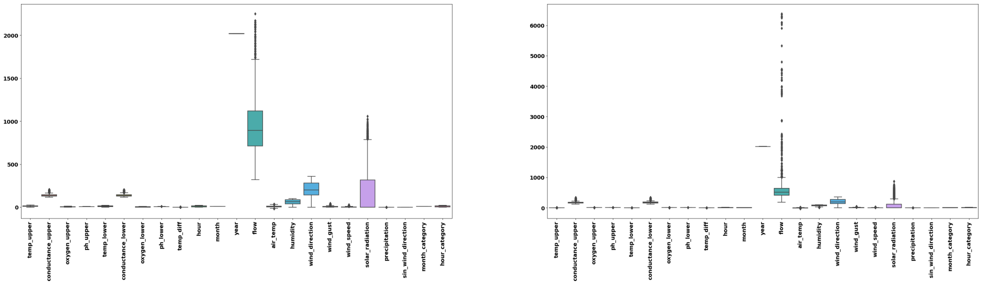


Figure 3: boxplot for fall and Winter for Miller

Moreover, an analysis of the 2016-2021 period was conducted using the Agrimet weather data obtained from the USGS. This investigation revealed a sharp decline in average air temperature in the Klamath River area from 2018 to 2019, with 2019 registering as the coolest year for average air temperature in the region, based on the available data. Subsequently, the average air temperature began to rise at an accelerating rate, culminating in 2021 as the year with the highest average air temperature recorded in the provided data.

In Figure 5, a sharp increase in precipitation is observed, which exhibits a negative correlation, or an inverse relationship, with air temperature. This correlation is further confirmed in the heatmap (Figure 1) when examining humidity against air temperature.

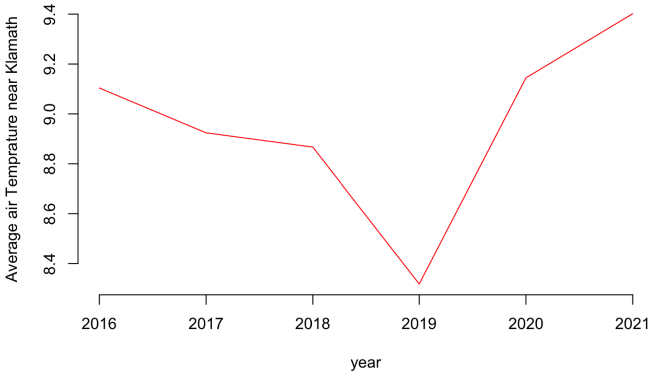


Figure 4: Yearly Air Temperature Average

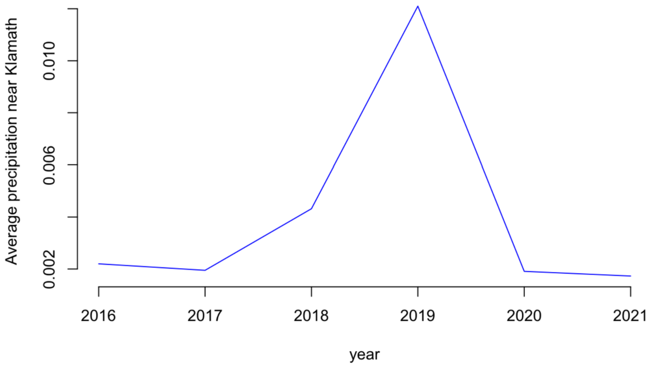


Figure 5: Yearly precipitation Average

0.5 Methods

0.5.1 Defining the Threshold for Thermal Stratification:

One of the primary objectives of this study is to establish a threshold temperature difference for thermal stratification in the Klamath River. The process begins by analyzing the distribution of the temperature difference variable within the dataset. Examining this distribution allows for the identification of a suitable threshold value that separates distinct groups or behaviors in the data. For instance, Figure 6 displays the distribution of temperature differences for the summer season. There are several general guidelines that can aid in selecting an appropriate threshold for temperature differences:

1. Identifying gaps or natural breaks in the histogram: If there is a clear gap between groups of bars in the histogram, this may be an ideal location for setting the threshold.

2. Determining the peak(s) in the histogram: If the histogram exhibits one or more peaks, consider placing the threshold between these peaks, particularly if the peaks represent different behaviors or phenomena.

3. Experimenting with various thresholds: Testing different thresholds and evaluating their effectiveness in separating the data into meaningful groups or categories

can provide valuable insights. Additional analyses, such as clustering or classification algorithms, can be employed to further refine the choice of threshold.

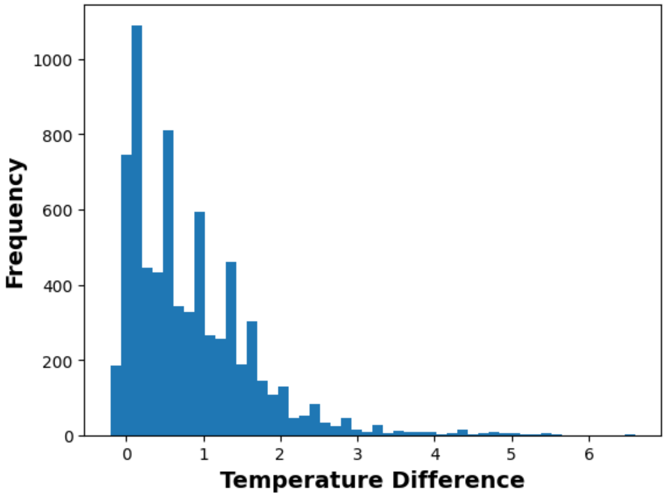


Figure 6: Miller Density plot for temperature difference in Summer

As depicted in Figure 6, the density plot for temperature differences in the summer season can be analyzed to identify potential thresholds. In addition to examining the distribution, clustering algorithms, such as K-means clustering, can be used to determine the choice of threshold. K-means is an unsupervised learning algorithm that aims to partition the data into K clusters. In this case, the goal is to divide the data into two clusters: stratified and unstratified. Since the objective is to find a threshold for temperature differences (temp-diff), only the 'temp-diff' column is used as a feature for K-means clustering. Once the cluster centroids are obtained, the threshold can be calculated as the midpoint between the two centroids, providing a well-defined criterion for thermal stratification.

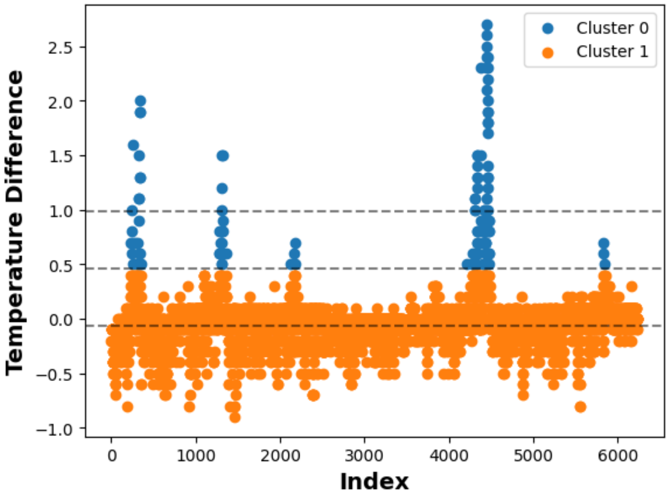


Figure 7: Miller Scatter plot of the 'temperature difference' cluster in Winter

0.5.2 Impact of weather conditions and river flow on the thermal stratification of the water column:

To study impact of weather conditions and water flow we explored Generalized additive models and Random forest. We will discuss why these two methods are a good fit in brief in this section.

GAM(generalized additive model):

A generalized additive model (GAM) is a statistical model that extends the capabilities of generalized linear models by allowing the estimation of complex nonlinear relationships between predictor variables and response variables. Unlike traditional linear models, GAMs do not assume a simple weighted sum relationship, but instead assume that the outcome can be modeled by a sum of arbitrary functions of each predictor feature. The relationship in a GAM is mathematically represented by a link function that establishes a relationship between the mean of the response variable and a smoothed function of the explanatory variable(s), which can be specified either parametrically or non-parametrically.

$$g(\mathrm{E}_Y(y|x)) = \beta_0 + \mathrm{f}_1(\mathrm{x}_1) + \mathrm{f}_2(\mathrm{x}_2) + ... + \mathrm{f}_p(\mathrm{x}_p) \tag{1}$$

Smooth functions in GAMs are represented as a sum of smaller basis functions, which capture nonlinear or time-dependent features of the data. GAMs can incorporate categorical variables, which allows for modeling of interactions between continuous and categorical predictors. GAMs have some limitations, including the potential to overfit the data, challenges in selecting appropriate basis functions, and computational expenses. Violations of model assumptions can also lead to biased or inefficient estimates. However, GAMs remain a useful tool for modeling nonlinear relationships between predictors and response variables in a flexible and interpretable manner.
Generalized additive mixed model(GAMMS) is an extension of GAM which gives a flexibility to work with correlated errors by defining suitable variance-covariance matrix.

0.5.3 Random forest:

Random Forest is a versatile algorithm for classification and regression tasks that combines multiple decision trees to create a more accurate and robust model. Known for its ability to handle complex datasets and its resilience to noise and outliers, it requires minimal feature engineering. The algorithm reduces overfitting and improves generalization by training decision trees on bootstrapped subsets of data and aggregating their predictions. Random forest is also a highly effective model for managing non-linear data, as well as addressing the inherent multicollinearity often present when working with weather and temporal data. The Random forest model can be mathematically represented as: $y_{RF} = \frac{1}{M}\sum_{m=1}^{M} y_m(x)$, where y_{RF} is the predicted output of the Random Forest model, M is the number of decision trees in the ensemble, $y_m(x)$ is the predicted output of the mth tree for input x, and the summation is taken over all trees in the ensemble.(3) The randomForest model in R provides several metrics that can be used to assess the success of a given Random Forest model and to gain insight on how important each predictor is for the model(4). The four main metrics provided by this library are outlined below:

1. Var explained: Derived from the Mean Squared Error (MSE) and the out-of-bag (OOB) predictions in the Random Forest model, a higher value for this metric indicates a better fit of the model to the data.
2. Mean of squared residuals: This metric represents the squared differences between the observed values and the predicted values for the response variable. A lower value indicates a better fit of the model to the data, as the model's predictions are closer to the actual values.
3. IncMSE (Percentage Increase in Mean Squared Error): This metric demonstrates the importance of a factor in the model. It is calculated by altering a factor's values to assess how much it throws off the predictions. A higher value means the factor is more crucial for accurate predictions.
4. IncNodePurity (Increase in Node Purity): This metric emphasizes the importance of each factor for precise classifications and predictions. A higher value indicates that the factor plays a more prominent role in enhancing the model's predictive performance.

0.5.4 Detecting Reverse Stratification Starting Date:

We have developed an R function that streamlines the process of identifying reverse stratification occurrences. This function performs the following steps:

1. Accepts a merged data frame and a span length in days as inputs.
2. Decomposes the timestamp into individual components: date, year, month, and day.
3. Calculates daily averages for variables such as temperature difference, flow, air temperature, and wind speed, allowing for the addition or removal of predictors as needed.
4. Computes the average of these variables over the specified span, with a default two-week period as an example.
5. Identifies the day in each year when the span-average temperature difference transitions from positive to negative.
6. Returns a data frame containing these dates, along with the span-average values for temperature difference, flow, air temperature, and wind speed, enabling the analysis of weather and flow conditions at the time of the change.

This method yields results indicating the specific day when the change occurs, as well as the average weather and flow conditions leading up to the event. To identify when thermal stratification reverts to positive, a similar approach can be employed, with the only modification being in step 5, where the transition from negative to positive is detected instead of positive to negative.

0.5.5 Detecting Intermittent Stratification:

Utilizing the provided merged dataset, we can employ the R functions group_by and summarise to aggregate the data by specific columns, such as hour, and subsequently compute the mean temperature difference. This process returns a data frame featuring each unique hour present in the dataset, along with the corresponding average temperature difference calculated over the entire dataset. Additionally, incorporating season or month alongside the hour allows for the examination of these averages across different seasons or months, depending on the intended analysis objectives. The hours with largest averages indicates the hours with greatest stratification. The hours with averages closest to zero indicate the hours that are most unstratified.

0.6 Result

GAM: In a GAM analysis, we examined the influence of time, air temperature, water flow, wind speed, wind direction, and precipitation on thermal stratification. We established a response variable as the difference in water temperature recorded by two sensors simultaneously.

We defined a variable 'weekday,' which represents the day of the week. As indicated in the literature, thermal stratification can be impacted by the quantity of dissolved particles or pollutants. We hypothesize that this could exhibit seasonality within weekdays. Each year may have a random effect on thermal stratification due to unique weather conditions specific to that year. We utilized a nonlinear temporal term to model the interaction of month, weekday, and hour, as there are seasonal effects influencing the EUI, such as weather or a higher quantity of pollutants. We established a non-linear interaction for air temperature (Figure 12), water flow (figure 13), precipitation, wind speed, and wind direction, as these features exhibit seasonal variations with the time of year and day.

Figure 14 demonstrates that the model assumes a Gaussian or normal distribution of errors, and the "Link" of "identity" indicates that the model does not transform the predictions. The effective degrees of freedom (edf) must be calculated to approximate the actual degrees of freedom for inference purposes. The edf represents the complexity of the smooth, with higher edfs describing more "wiggly" curves. The Ref.df and F columns are used in an ANOVA test to assess the overall significance of the smooth, with the p-value indicating the result of this test. The GAM's smooth terms are plotted in figures 15, 16, 17, 18, 19, 20, and 21 illustrating the seasonality of water temperature difference over the interaction of month, hour, and weekday. GAM plots for water flow (Figure 18) and air temperature (Figure 19) reveal that flow and air temperature have a non-linear relationship with water temperature difference. Figure 15 shows that the stratification is significant between late spring and early fall during early morning hours and late evening. It also shows significant reverse stratification between late fall and early winters .

To ensure well-fit models, it is necessary to avoid several pitfalls when fitting GAMs, such as ensuring the correct number of basis functions to account for wiggly data and normality assumptions. Figure 22 displays the output from gam.check(), which reports on model convergence, demonstrating full convergence in this case. The table of basis checking results indicates a statistical test for patterns in model residuals, which should

be random. P-values greater than the level of significance suggest that residuals are not randomly distributed, indicating that the number of basis functions needs to be increased.
Furthermore, we must verify if concurvity exists between our smoothed variables, which refers to linear dependence between smooths. Figure 21 exhibits the concurvity, with values higher than the conventional value of 0.8 indicating the presence of concurvity and potentially inaccurate results.
Lastly, the outcomes of Figure 23 present the standard checks for the normality assumption of the error terms. If the normality assumptions are not met, potential fixes include increasing the number of basis functions, altering the link function, or applying a transformation on the response, such as a log transformation.

Random Forest: In this example analysis, as shown in Figure 25, a Random Forest model was implemented using the randomForest library and function in R. The response variable, temperature difference, was modeled as a function of several predictor variables, including flow, cardinal wind direction, air temperature, wind speed, month, and hour. The model was trained on the keno dataset, with an ensemble of 501 decision trees and the importance of each variable set to true.

To evaluate the model's performance, we refer to the metrics outlined in the methods section. Firstly, a percentage of variance explained of 87.68% suggests that the model accounts for a substantial portion of the variance in the response variable, indicating a good fit to the data. A mean squared residuals value of 0.072 demonstrates that the model is accurate in making predictions, as it measures the squared differences between the observed and predicted values of the response variable.

The other metrics assess the importance of each predictor variable, as shown in Figure 8 below. The hour predictor ranks second in %IncMSE and third in IncNodePurity, implying its importance in the model. In contrast, the cardinal wind direction has relatively lower %IncMSE and IncNodePurity values, indicating it is less significant compared to the other predictors. Air temperature has the highest %IncMSE and IncNodePurity, signifying it as the most critical predictor in the model. Wind speed exhibits the lowest %IncMSE and a relatively lower IncNodePurity, suggesting it is the least important predictor among the variables. The month variable holds a moderate %IncMSE and the second-highest IncNodePurity, implying its importance in the model. The flow variable has third highest %IncMSE and third lowest IncNodePurity values, indicating its in the middle of significance in the model.

In summary, this Random Forest model demonstrates strong performance with an explained variance of 87.68%. Air temperature emerges as the most important predictor, followed by flow and hour. Month also holds considerable importance, while cardinal wind direction and wind speed are less influential predictors in the model.

Reverse Stratification: The example analysis used the R function provided, taking in the merged Keno dataset and a span of 14 days. As shown in figure 26, the starting

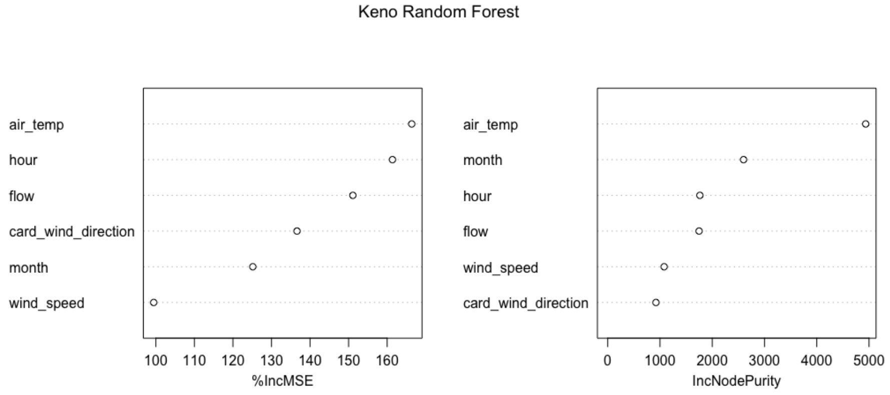


Figure 8: Keno Random Forest Predictor Importance

date for reverse stratification occurred as early as 11/19 (2020) and as late as 12/07 (2017). The results from reverse stratification back to stratification have it occurring as early as 01/25 (2017) to as late as 03/06 (2019)

Intermittent Stratification: For this example, as shown in Figure 27, only hours and averages are shown. The hours with highest temperature difference are between 4pm and 6pm, and the hours closest to zero are between 7am and 9am.

0.7 Recommendations

Cleaning and merging data: We suggest cleaning each of the four data sets prior to merging them. Many water quality data points were recorded at the 59th minute of the hour, so it is advisable to round timestamps to the nearest hour within a five-minute window to retain these data points. After rounding, the data sets can be merged using exact timestamps. By consolidating all response and predictor data into a single data frame, the process of running models in R is simplified. New columns can be created to analyze seasonal, monthly, and hourly variations. We recommend transforming wind direction into cardinal wind directions or applying the sine function to account for its cyclical nature. Furthermore, it is advisable to convert seasonal, monthly, hourly, and cardinal wind directions into factors, as they possess a categorical nature.

Running models: When implementing GAM and Random Forest models, it is crucial to avoid incorporating predictor variables that exhibit high correlations with one another to prevent unnecessary multicollinearity. We suggest examining the provided correlation matrix for predictors and avoiding combining those with higher correlations with the other predictors. For instance, it is not advisable to combine solar radiation, air temperature, or humidity within the same model. Instead, test each one independently, excluding the other two, and assess which variable produces the best results based on metrics such as adjusted R-squared value for GAM or the percentage of variance explained for Random Forest.

To gain a deeper understanding of the influence each predictor has on the response variable, individual models can be run with a single predictor at a time. The more significant a predictor's impact on the response variable, the higher the adjusted R-squared value or explained variance. A common approach for developing a comprehensive statistical model involves what is called forward selection, which is a type of stepwise regression (5). In forward selection, predictors are added one at a time to the model, starting with the predictor that has the highest adjusted R-squared value. At each step, the predictor that results in the largest increase in adjusted R-squared is chose, and the process is repeated until no further significant improvement in adjusted R-squared is observed or there are no more predictors to add.

0.8 Conclusion

In conclusion, traditional regression techniques are often inadequate for dealing with time series data due to autocorrelation and their limitation to linear relationships. Instead, Generalized Additive Models (GAMs) have emerged as a viable alternative for modeling nonlinear relationships and capturing complex patterns in time series data, much like Random Forest. Additionally, we recommend analyzing the distribution of the temp-diff variable and employing K-means clustering can aid in defining an appropriate threshold for thermal stratification.

0.9 References

1: US Geological Survey. (n.d.). Keno Reach Study Area Map.
2: U.S. Geological Survey. (2021, April 22). Buoy Platforms for Monitoring Water Quality Deployed. USGS.gov.
3: Simpson, G. Modelling Seasonal Data with GAM. From the Bottome of the Heap.(2014, May 9).
4: Breiman, L. Random Forests. Machine Learning 45, 5-32 (2001).
5: Liaw, A and Wiener, M. Classification and Regression by randomForest. R News 2(3) 18-22 (2002)
6: Draper, N.R. and Smith, H. Applied Regression. 3rd Edition, Wiley, New York. 327-368 (1998)
7: Hartigan, J. A., Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Journal of the royal statistical society. series c (applied statistics), 28(1), 100-108.

0.10 Appendix

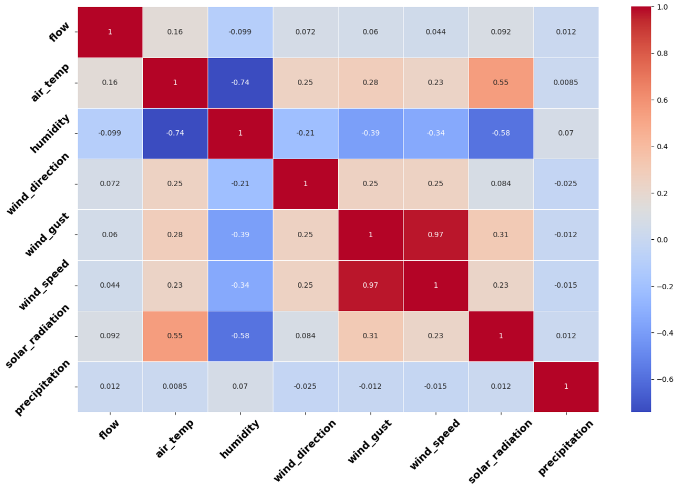


Figure 9: Correlation plot between predictors for Keno

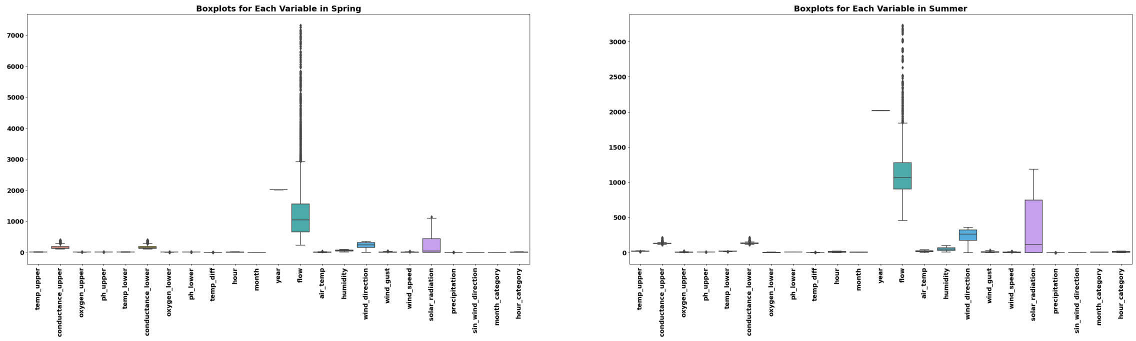


Figure 10: boxplot for Spring and Summer for Miller

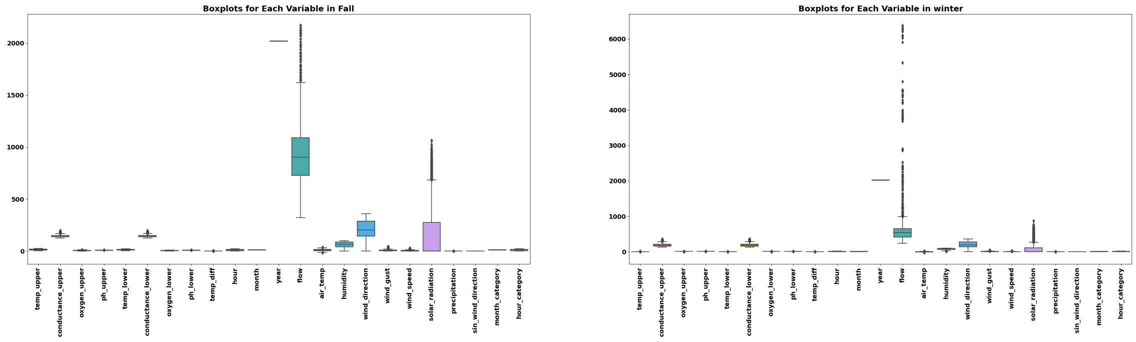


Figure 11: boxplot for fall and Winter for Miller

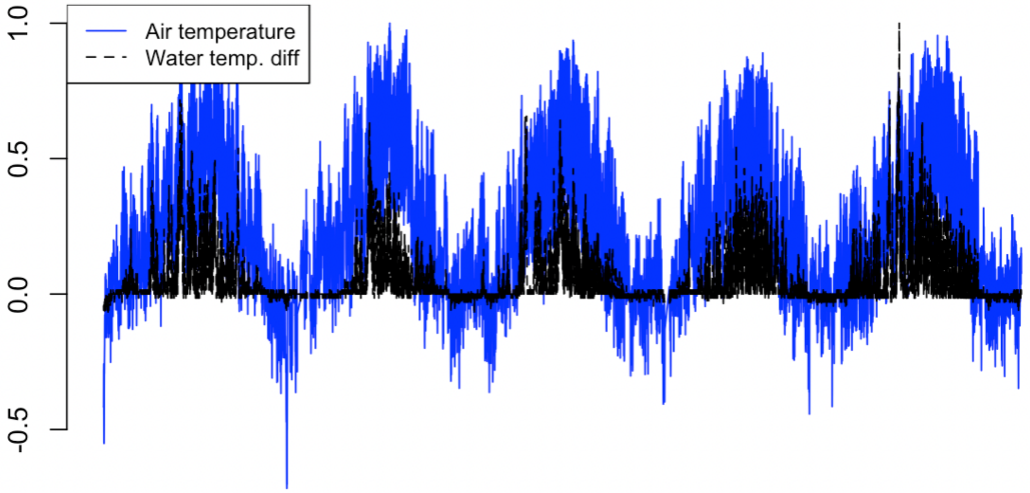


Figure 12: Air-temperature and Water temperature difference for training data

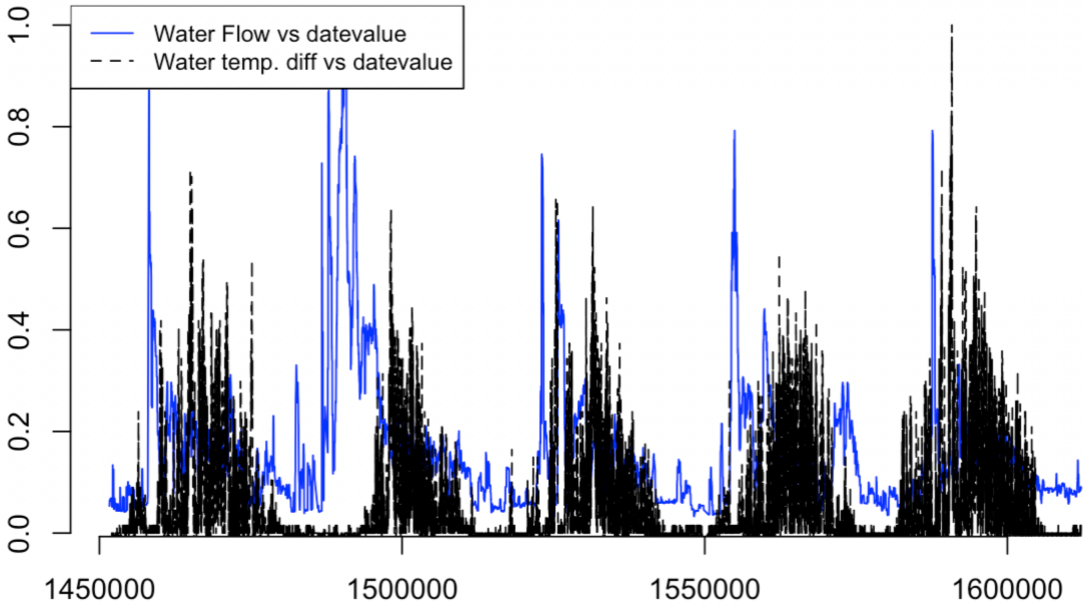
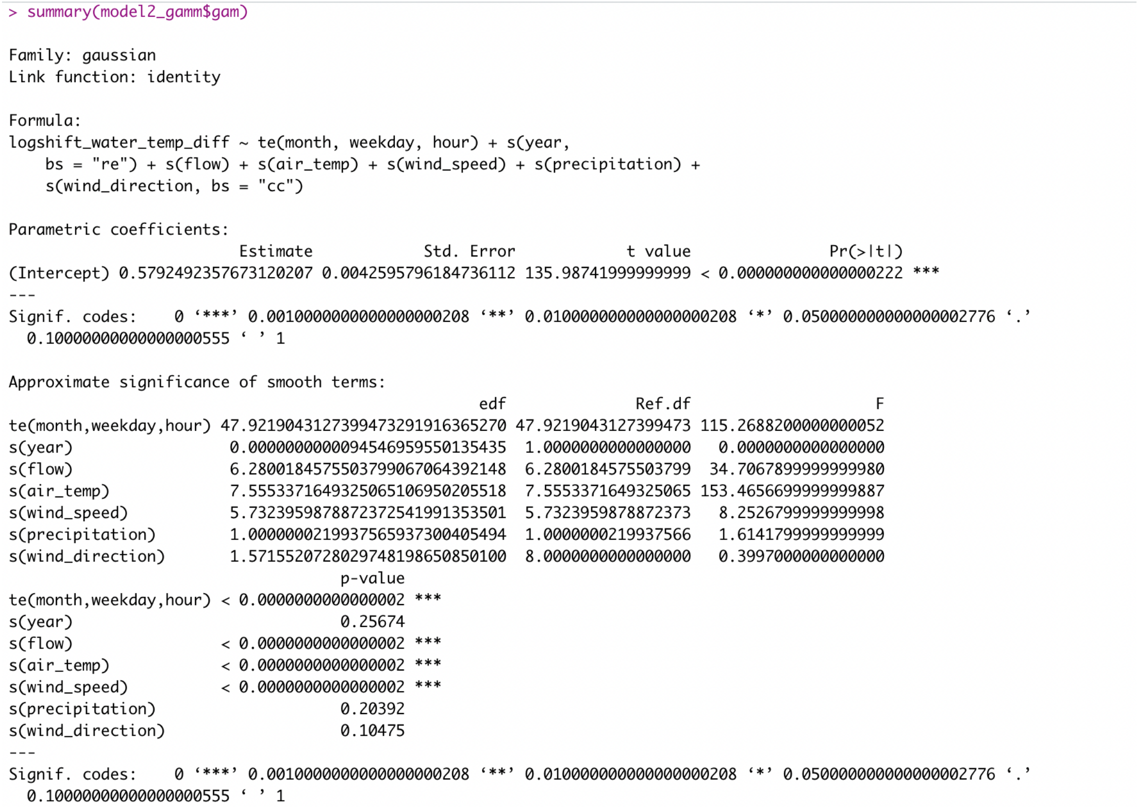


Figure 13: Water flow and Water temperature difference for training data



```
> summary(model2_gamm$gam)

Family: gaussian
Link function: identity

Formula:
logshift_water_temp_diff ~ te(month, weekday, hour) + s(year,
    bs = "re") + s(flow) + s(air_temp) + s(wind_speed) + s(precipitation) +
    s(wind_direction, bs = "cc")

Parametric coefficients:
                         Estimate            Std. Error            t value               Pr(>|t|)
(Intercept) 0.5792492357673120207 0.0042595796184736112 135.98741999999999 < 0.0000000000000000222 ***
---
Signif. codes:    0 ‘***’ 0.0010000000000000000208 ‘**’ 0.0100000000000000000208 ‘*’ 0.0500000000000000002776 ‘.’
  0.10000000000000000555 ‘ ’ 1

Approximate significance of smooth terms:
                                                edf               Ref.df                     F
te(month,weekday,hour) 47.9219043127399473291916365270 47.9219043127399473 115.2688200000000052
s(year)                 0.00000000000094546959550135435  1.0000000000000000   0.0000000000000000
s(flow)                 6.2800184575503799067064392148  6.2800184575503799  34.7067899999999980
s(air_temp)             7.5553371649325065106950205518  7.5553371649325065 153.4656699999999887
s(wind_speed)           5.7323959878872372541991353501  5.7323959878872373   8.2526799999999998
s(precipitation)        1.0000000219937565937300405494  1.0000000219937566   1.6141799999999999
s(wind_direction)       1.5715520728029748198650850100  8.0000000000000000   0.3997000000000000
                                  p-value
te(month,weekday,hour) < 0.0000000000000002 ***
s(year)                             0.25674
s(flow)                < 0.0000000000000002 ***
s(air_temp)            < 0.0000000000000002 ***
s(wind_speed)          < 0.0000000000000002 ***
s(precipitation)                    0.20392
s(wind_direction)                   0.10475
---
Signif. codes:    0 ‘***’ 0.0010000000000000000208 ‘**’ 0.0100000000000000000208 ‘*’ 0.0500000000000000002776 ‘.’
  0.10000000000000000555 ‘ ’ 1
```

Figure 14: Output from summary of GAM

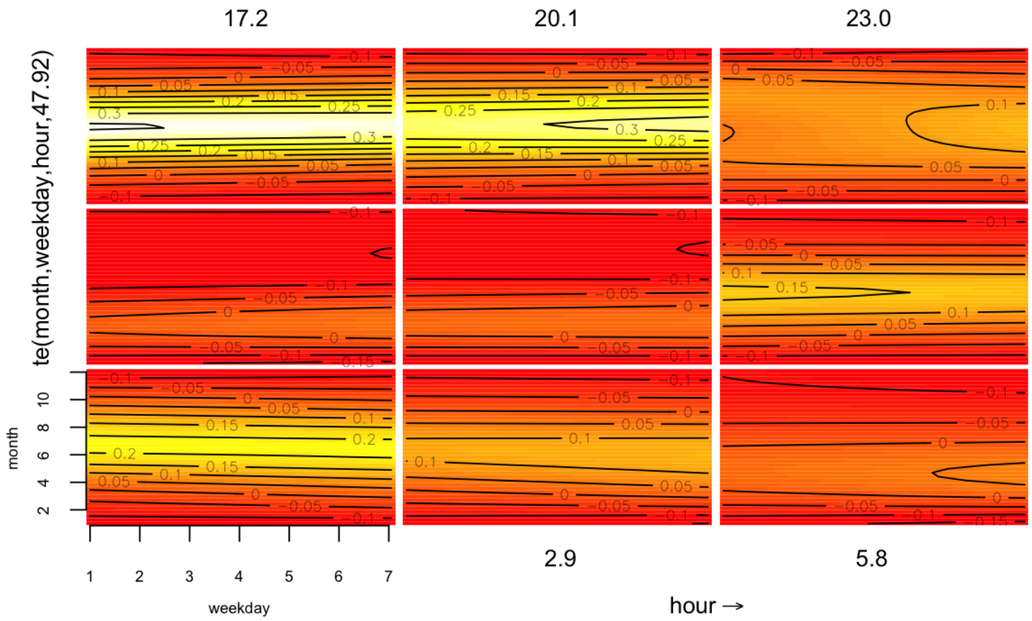


Figure 15: Effect of (month, weekday, hour) interactions on water temperature difference at bottom and surface layers

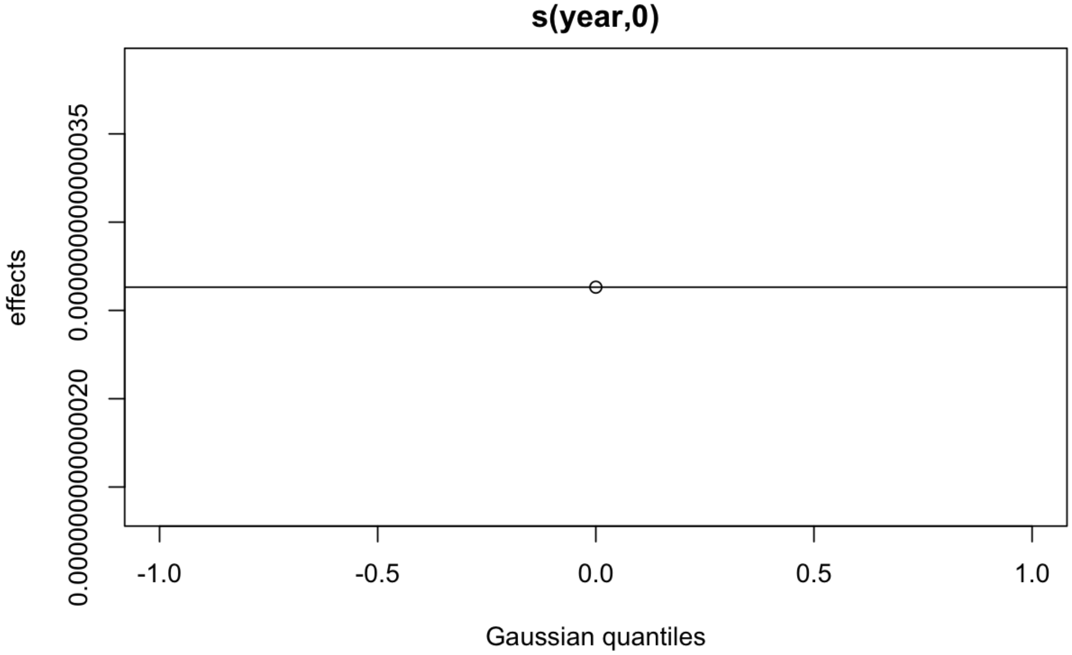


Figure 16: Effect of year on water temperature difference at bottom and surface layers

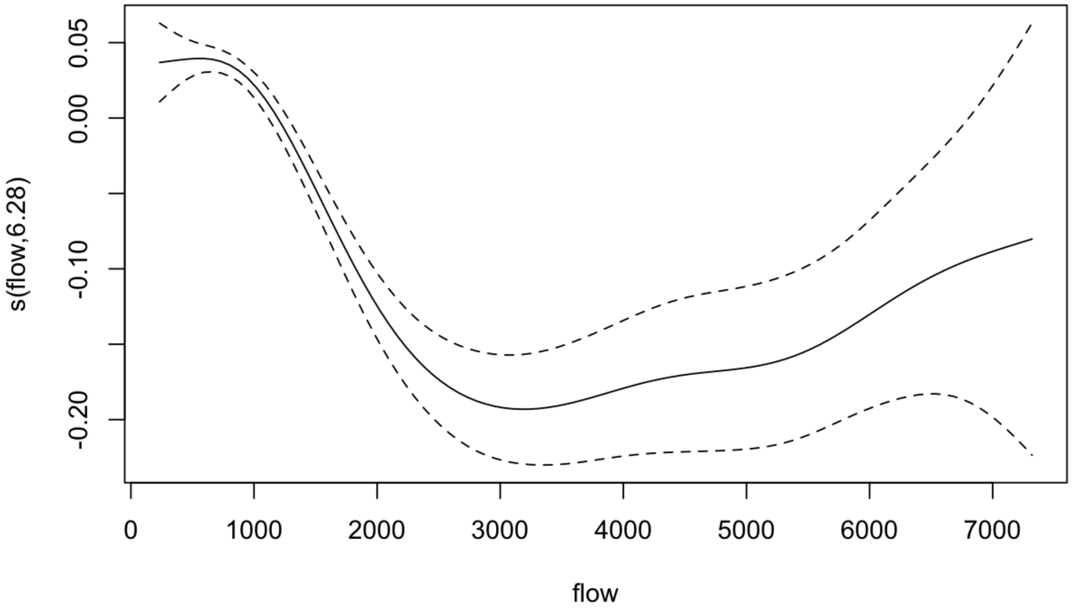


Figure 17: Effect water flow on water temperature difference at bottom and surface layers

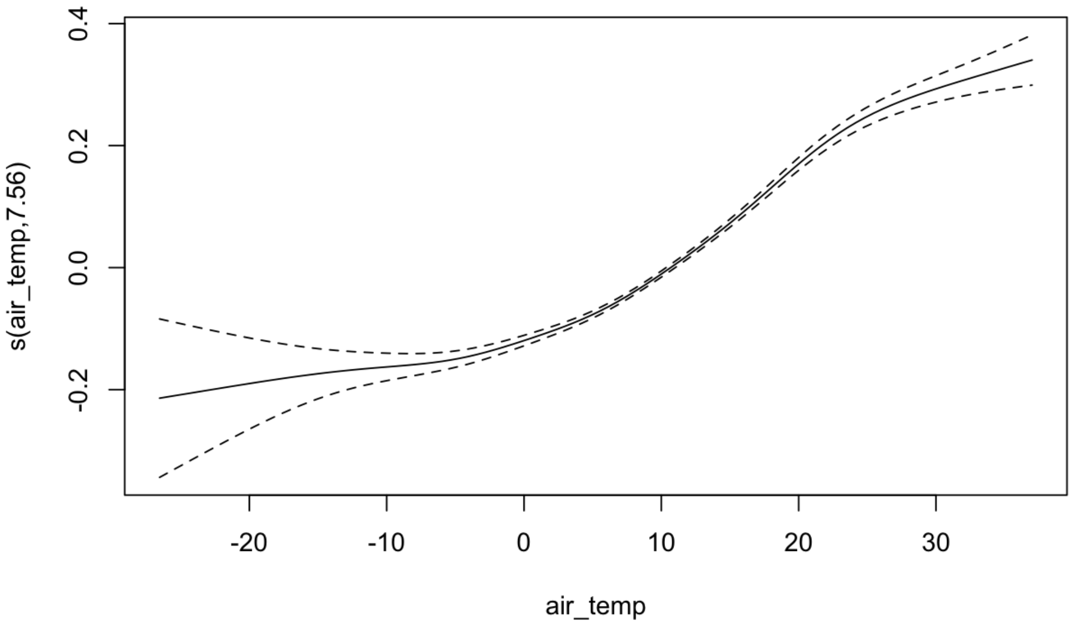


Figure 18: Effect air temperature on water temperature difference at bottom and surface layers

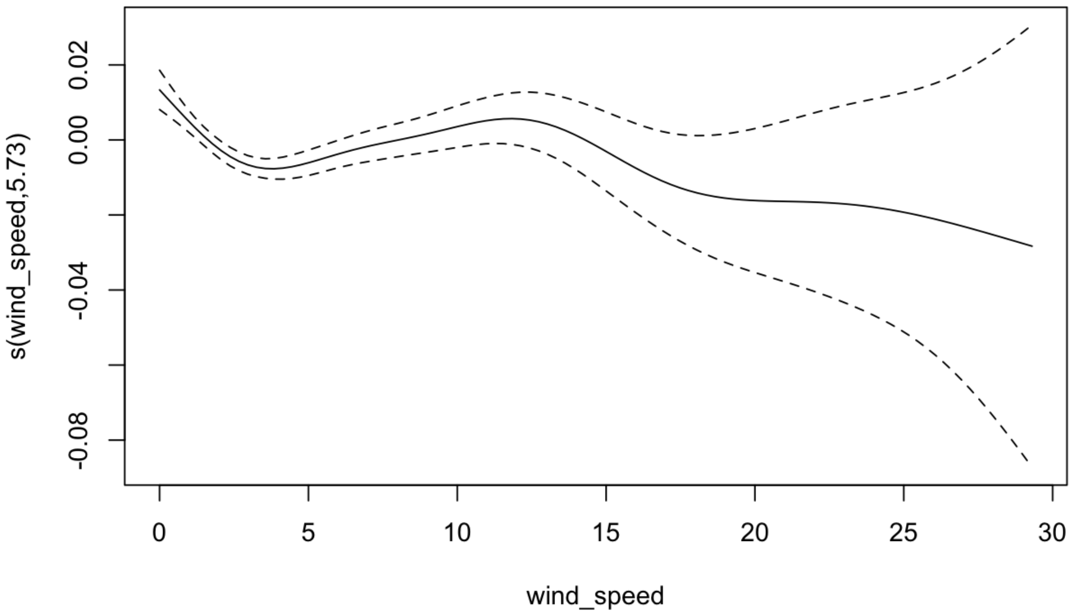


Figure 19: Effect wind speed on water temperature difference at bottom and surface layers

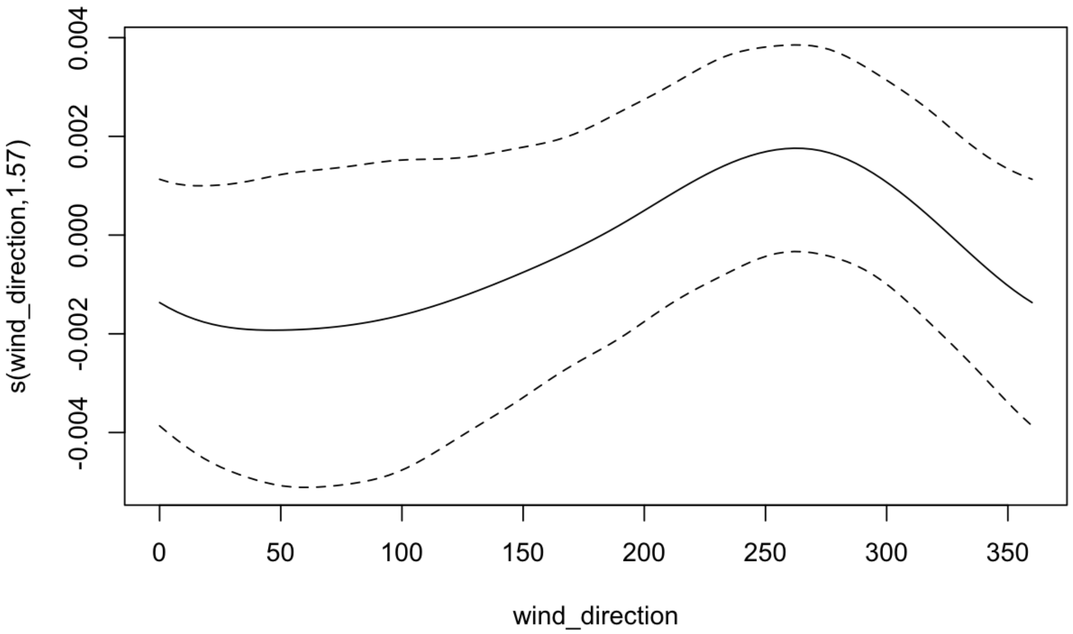


Figure 20: Effect wind direction on water temperature difference at bottom and surface layers

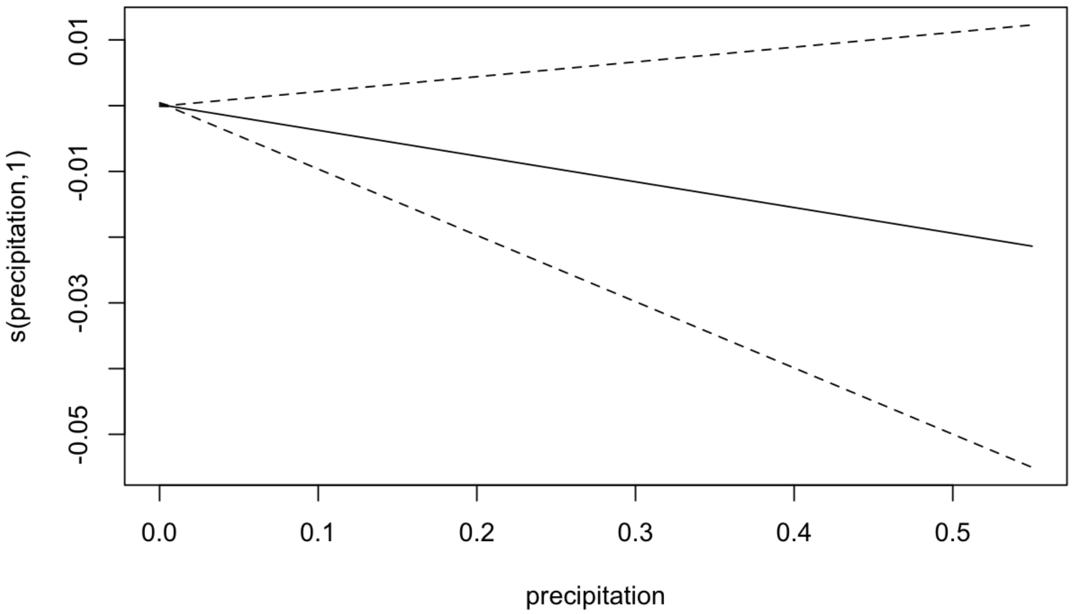


Figure 21: Effect precipitation on water temperature difference at bottom and surface layers

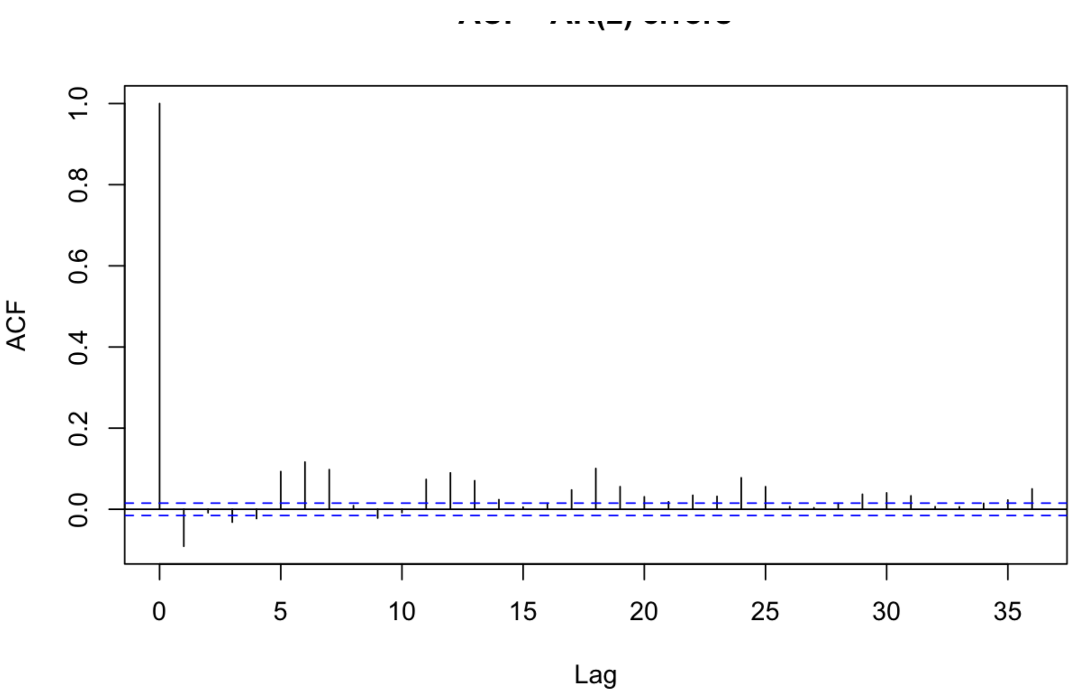
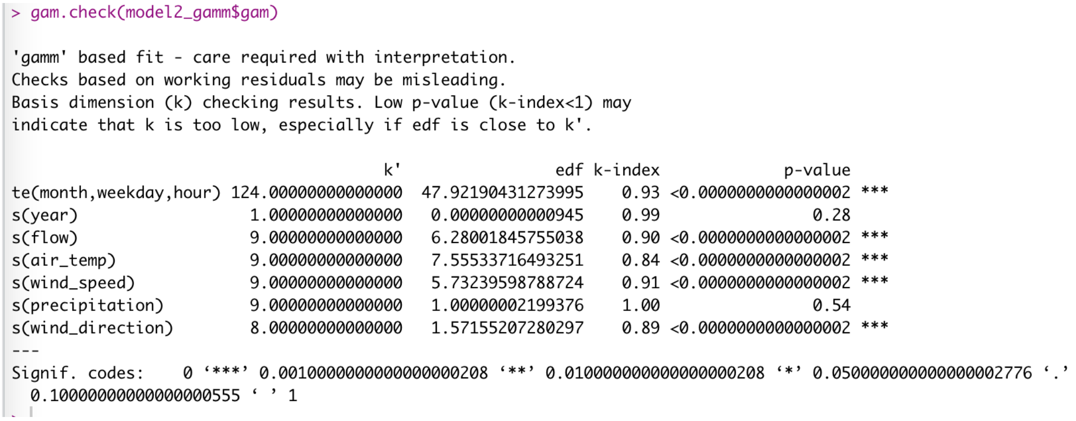
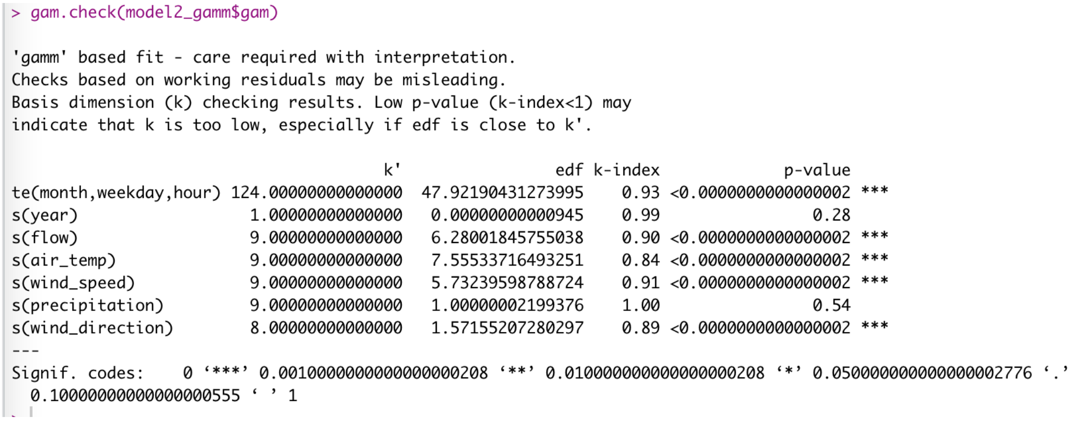


Figure 22: Checking for autocorrelation



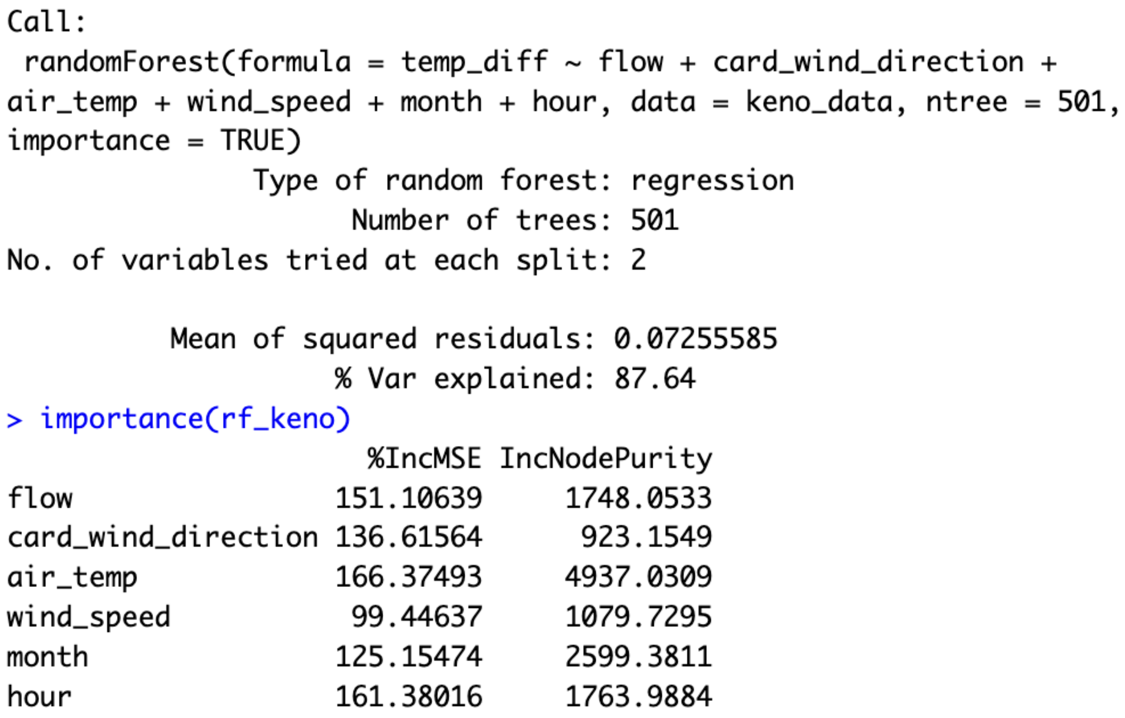
```
> gam.check(model2_gamm$gam)

'gamm' based fit - care required with interpretation.
Checks based on working residuals may be misleading.
Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                                     k'                edf k-index              p-value    
te(month,weekday,hour) 124.00000000000000  47.92190431273995    0.93 <0.0000000000000002 ***
s(year)                  1.00000000000000   0.00000000000945    0.99                0.28    
s(flow)                  9.00000000000000   6.28001845755038    0.90 <0.0000000000000002 ***
s(air_temp)              9.00000000000000   7.55533716493251    0.84 <0.0000000000000002 ***
s(wind_speed)            9.00000000000000   5.73239598788724    0.91 <0.0000000000000002 ***
s(precipitation)         9.00000000000000   1.00000002199376    1.00                0.54    
s(wind_direction)        8.00000000000000   1.57155207280297    0.89 <0.0000000000000002 ***
---
Signif. codes:    0 ‘***’ 0.00100000000000000002 ‘**’ 0.01000000000000000020 ‘*’ 0.05000000000000000277 ‘.’ 
  0.10000000000000000555 ‘ ’ 1
```

Figure 23: Concurvity of model variables



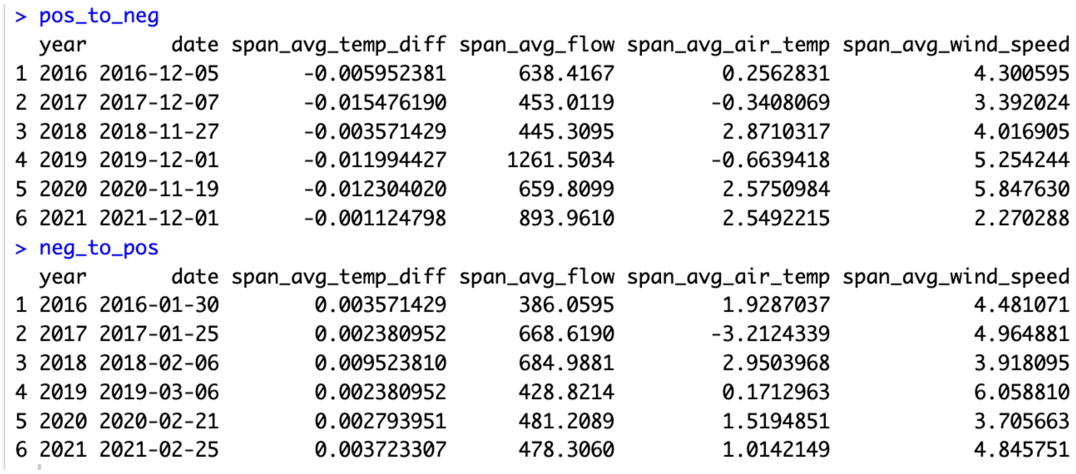
```
> gam.check(model2_gamm$gam)

'gamm' based fit - care required with interpretation.
Checks based on working residuals may be misleading.
Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                                     k'                edf k-index              p-value    
te(month,weekday,hour) 124.00000000000000  47.92190431273995    0.93 <0.0000000000000002 ***
s(year)                  1.00000000000000   0.00000000000945    0.99                0.28    
s(flow)                  9.00000000000000   6.28001845755038    0.90 <0.0000000000000002 ***
s(air_temp)              9.00000000000000   7.55533716493251    0.84 <0.0000000000000002 ***
s(wind_speed)            9.00000000000000   5.73239598788724    0.91 <0.0000000000000002 ***
s(precipitation)         9.00000000000000   1.00000002199376    1.00                0.54    
s(wind_direction)        8.00000000000000   1.57155207280297    0.89 <0.0000000000000002 ***
---
Signif. codes:    0 ‘***’ 0.00100000000000000002 ‘**’ 0.01000000000000000020 ‘*’ 0.05000000000000000277 ‘.’ 
  0.10000000000000000555 ‘ ’ 1
```

Figure 24: Output from gam.check()



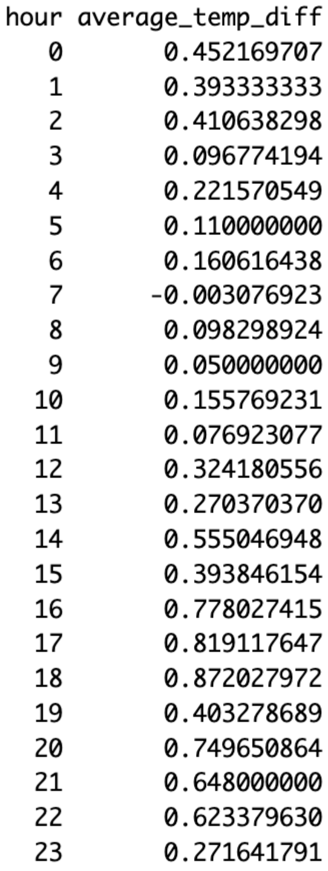
```
Call:
 randomForest(formula = temp_diff ~ flow + card_wind_direction +
air_temp + wind_speed + month + hour, data = keno_data, ntree = 501,
importance = TRUE)
               Type of random forest: regression
                     Number of trees: 501
No. of variables tried at each split: 2

          Mean of squared residuals: 0.07255585
                    % Var explained: 87.64
> importance(rf_keno)
                      %IncMSE IncNodePurity
flow                151.10639     1748.0533
card_wind_direction 136.61564      923.1549
air_temp            166.37493     4937.0309
wind_speed           99.44637     1079.7295
month               125.15474     2599.3811
hour                161.38016     1763.9884
```

Figure 25: Output from Random Forest



```
> pos_to_neg
  year       date span_avg_temp_diff span_avg_flow span_avg_air_temp span_avg_wind_speed
1 2016 2016-12-05       -0.005952381      638.4167         0.2562831            4.300595
2 2017 2017-12-07       -0.015476190      453.0119        -0.3408069            3.392024
3 2018 2018-11-27       -0.003571429      445.3095         2.8710317            4.016905
4 2019 2019-12-01       -0.011994427     1261.5034        -0.6639418            5.254244
5 2020 2020-11-19       -0.012304020      659.8099         2.5750984            5.847630
6 2021 2021-12-01       -0.001124798      893.9610         2.5492215            2.270288
> neg_to_pos
  year       date span_avg_temp_diff span_avg_flow span_avg_air_temp span_avg_wind_speed
1 2016 2016-01-30        0.003571429      386.0595         1.9287037            4.481071
2 2017 2017-01-25        0.002380952      668.6190        -3.2124339            4.964881
3 2018 2018-02-06        0.009523810      684.9881         2.9503968            3.918095
4 2019 2019-03-06        0.002380952      428.8214         0.1712963            6.058810
5 2020 2020-02-21        0.002793951      481.2089         1.5194851            3.705663
6 2021 2021-02-25        0.003723307      478.3060         1.0142149            4.845751
```

Figure 26: Output date changes for reverse stratification



```
hour average_temp_diff
   0       0.452169707
   1       0.393333333
   2       0.410638298
   3       0.096774194
   4       0.221570549
   5       0.110000000
   6       0.160616438
   7      -0.003076923
   8       0.098298924
   9       0.050000000
  10       0.155769231
  11       0.076923077
  12       0.324180556
  13       0.270370370
  14       0.555046948
  15       0.393846154
  16       0.778027415
  17       0.819117647
  18       0.872027972
  19       0.403278689
  20       0.749650864
  21       0.648000000
  22       0.623379630
  23       0.271641791
```

Figure 27: Output hours and average temperatures