Prediction of Drinking Water Treatment Process Chemical Doses

Dawson Carney, Reece Carmody, Ethan Schilling, Joshua Tobey

17 December 2025

Contents

Introduction **2**
- Background and Motivation . . . 2
- Project Purpose . . . 2

Data Overview and Cleaning **3**
- Data Overview . . . 3
- Data Cleaning . . . 3
- Exploratory Analysis . . . 4
  - Alkalinity and Hardness Data . . . 4
  - Primary Raw Water Characteristic Data . . . 4
  - Coagulant Dose Data . . . 5
- Predictor Selection . . . 6
- Added Variables . . . 6

Modeling **6**
- Summary of Modeling Efforts . . . 6
  - Modeling Priorities . . . 6
  - Training and Testing Datasets . . . 6
  - Models Used . . . 7
- Dose Prediction Models . . . 7
  - Overview of Dose Prediction . . . 7
  - Linear Regression . . . 8
  - Decision Tree . . . 10
  - Random Forest . . . 11
  - Generalized Additive Model (GAM) . . . 12
  - Summary of Dose Prediction Models . . . 13
- Dose Change Prediction Models . . . 13
  - Overview of Dose Change Prediction . . . 13
  - Linear Discriminant Analysis (LDA) . . . 14
  - Quadratic Discriminant Analysis (QDA) . . . 14
  - Multinomial Logistic Regression . . . 15
  - Random Forest . . . 15
  - Summary of Dose Change Prediction Models . . . 15

Conclusion **16**

CAUTION: Please uncomment any of the install.packages that aren't already installed.

Introduction

Background and Motivation

The drinking water treatment process takes water from a river, lake, reservoir, or other source, and purifies it to have it reach drinking water standards.

There are four primary steps of the water treatment process:
1. Coagulation - Chemicals ("coagulants") are added to raw water to help contaminants group together into "floc" particles
2. Flocculation - The water is then slowly mixed to allow these floc particles to grow
3. Filtration - These particles are filtered out
4. Disinfection - Water is disinfected to get rid of biological contaminants

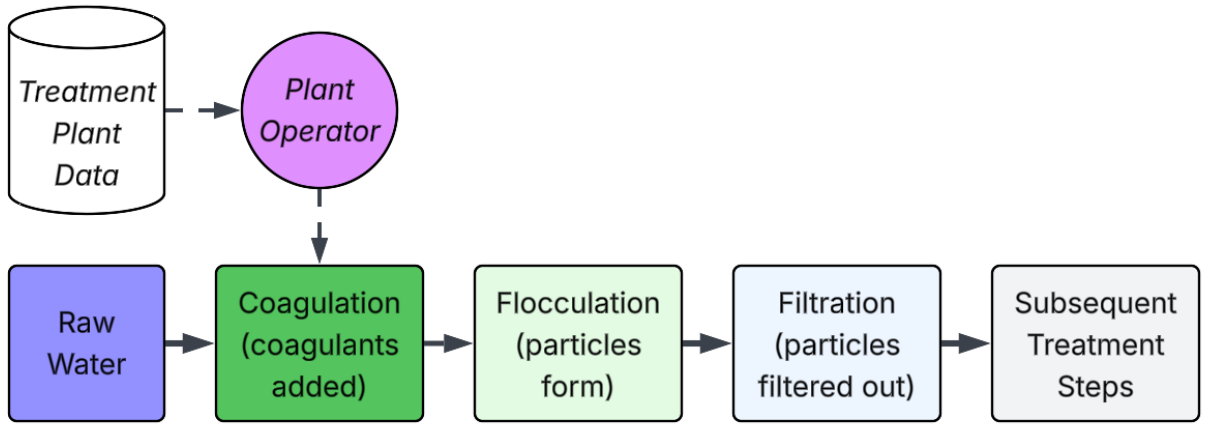
This project will focus on Step 1. These chemical doses are one of the most important features of the treatment process that an operator can change to improve performance.

Many different factors indicate effective chemical dose performance. Examples could include:

- chemical information about the water such as turbidity, charge, or pH
- information on the size of the particles being formed in the flocculation process (found via running tests on water samples)
- observation of filter performance at the end of the process
- total chemical byproducts produced
- output water quality

The chemical processes that determine treatment process performance are complex and highly interconnected, therefore all of these factors and more are important in making effective decisions about chemical dosing.

The role of an operator in deciding water treatment plant chemical doses is to look at a wide array of these factors and make decisions about how to adjust chemical doses. The treatment process and operator role are summarized in the figure below.



Treatment plant conditions can change rapidly, so having tools that aid in selecting chemical doses in response to changing water quality is crucial for operators. These chemicals are one of the primary costs of the treatment process, costing millions of dollars per year for a mid-size treatment plant. The chemical byproducts produced by excessive coagulant doses also have environmental impacts. Therefore, this is both a financial problem impacting taxpayers, and an environmental one.

Project Purpose

The goal of this project is to produce a model that serves the role of a treatment plant operator. This model would provide recommended chemical doses, based on input water quality characteristics, and

how operators have dosed chemicals in the past. This could prove a useful tool to operators as a "starting point" in chemical dosing. They could see a change in water quality, retrieve the suggested chemical dose from the model, test this dose, and adjust from there based on other information. We will attempt to create both a predictive model for predicting the chemical dose, and a model to predict dose change (increase, decrease, stay same), both based on raw water characteristics.

Data Overview and Cleaning

Data Overview

The data used for this effort is a timeseries dataset from a Colorado water treatment plant covering 3 years, from 2018-2020. This treatment plant takes its water from a reservoir, which tends to be a more stable water source than sources such as rivers or industrial supplies.

The available data are as follows:

- Raw Water Data
  - pH
  - Temperature
  - Turbidity ("cloudiness" of the water)
  - Suspended Grain Size Distributions
  - Alkalinity (resistance of water to changes in pH)
  - Hardness (mineral content of water)
- Chemical Dosing Data
  - Coagulant Dose - primary additive (allows for floc particle formation)
  - Cationic Polymer Dose - secondary additive (boosts size of floc particles)

Data Cleaning

Basic data cleaning was conducted based on visual inspection. Most of the raw water data we have available was on a 4-hour time increment. So, we decided to combine and average by day for our predictions. This is because the dosing data is only available as daily averages, so we needed to match our time resolution.

Before this averaging, we removed or modified values that were clearly outliers, so the averages would not be skewed by incorrect data. Reasons for these outliers are most likely equipment malfunction.

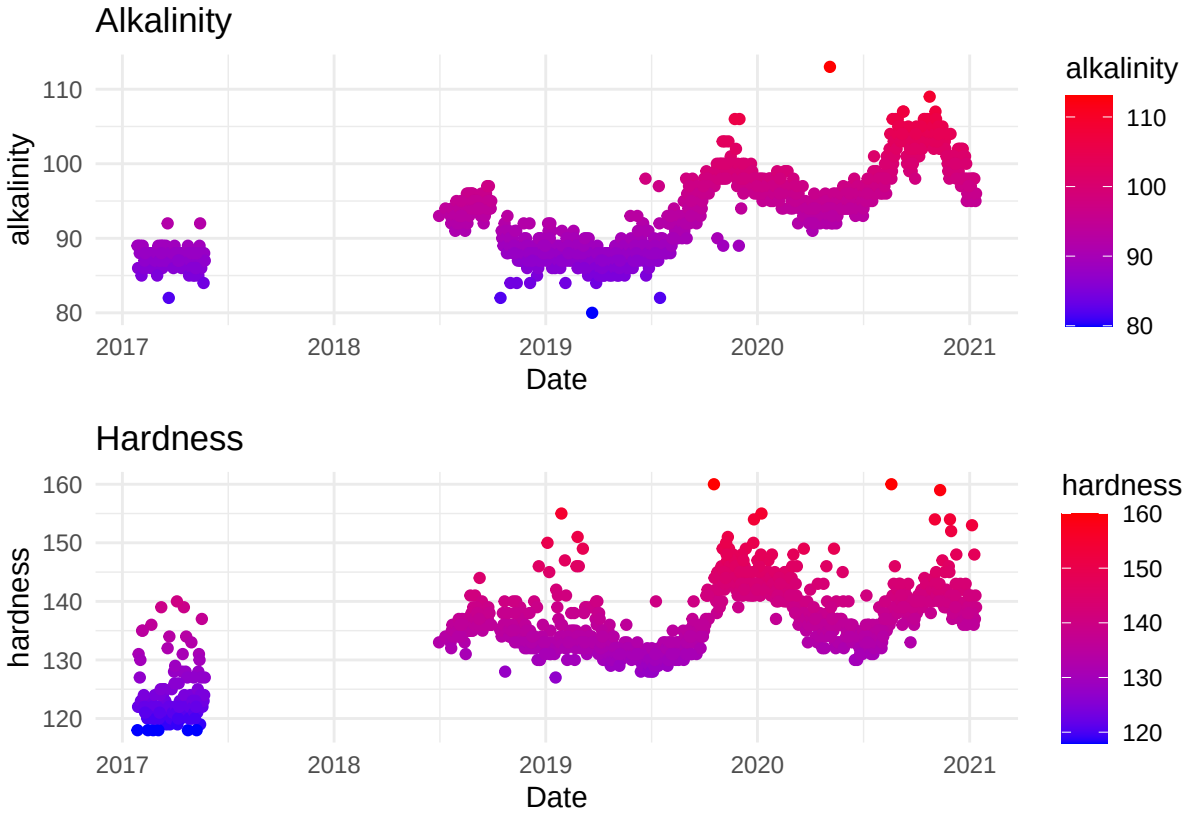
- **pH**: No significant outliers were present that we could see, so we simply day-averaged.
- **Temperature**: There was one region of approximately zero temperature (unrealistic for the reservoir). We got rid of these values and interpolated to the nearest non-zero temperature.
- **Conductivity**: Values less than 0.2 and greater than 2 were removed, since they are unreasonable based on this dataset and the "usual" values.
- **Turbidity**: A few exceptionally high values (relative to the usual data values) were removed before averaging.
- **Grain Size Information**: This data was highly variable, and simply looks like "noise", so did not seem helpful for prediction. We thus did not clean it.
- **Alkalinity and Hardness**: These data have different time availability than the other datasets but will be inspected. Removed zero-values.

Final Combination: We merged the raw water data (excluding alkalinity/hardness) with the dosing data. We only lost about 10 records in this process, where there was not available dosing data for the raw water measurements.

Exploratory Analysis

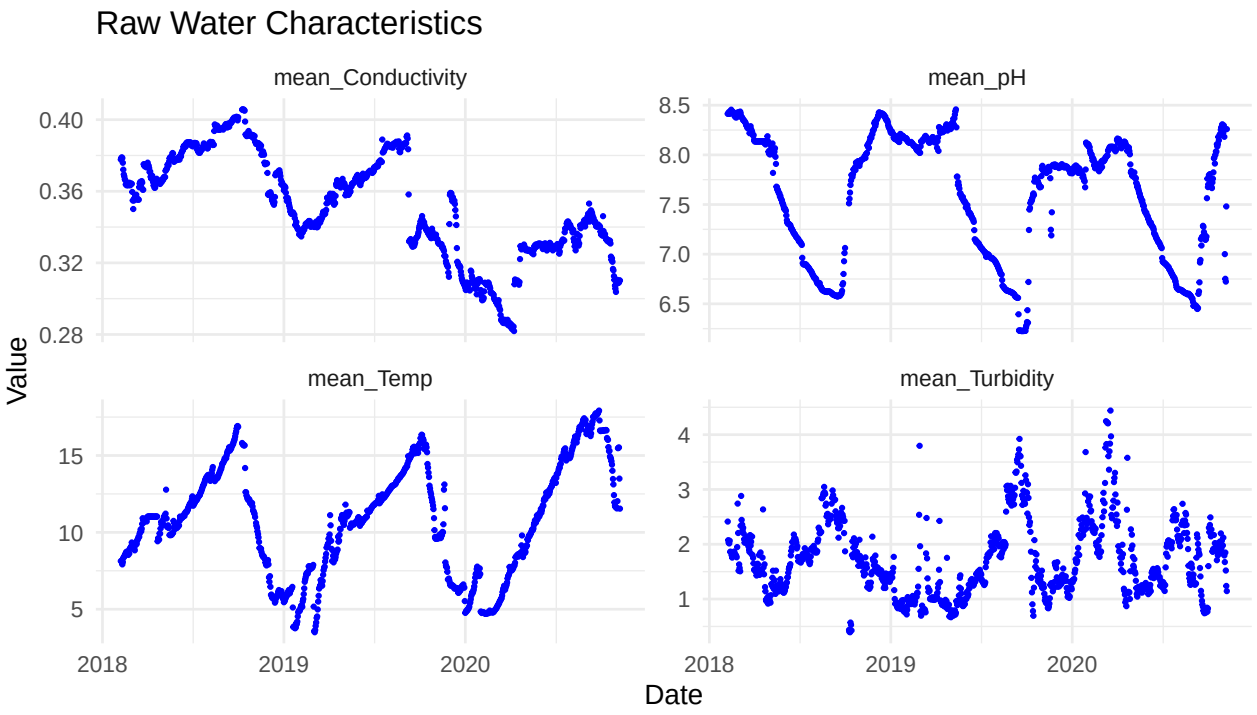
Alkalinity and Hardness Data

While alkalinity and hardness display seasonal trends that could be helpful for prediction, the available time ranges of the data did not align with other datasets (namely, dosing and other raw water characteristics), so they will not be used in this analysis. However, they are pictured below. Note the gap in data availability from mid-2017 to mid-2018.



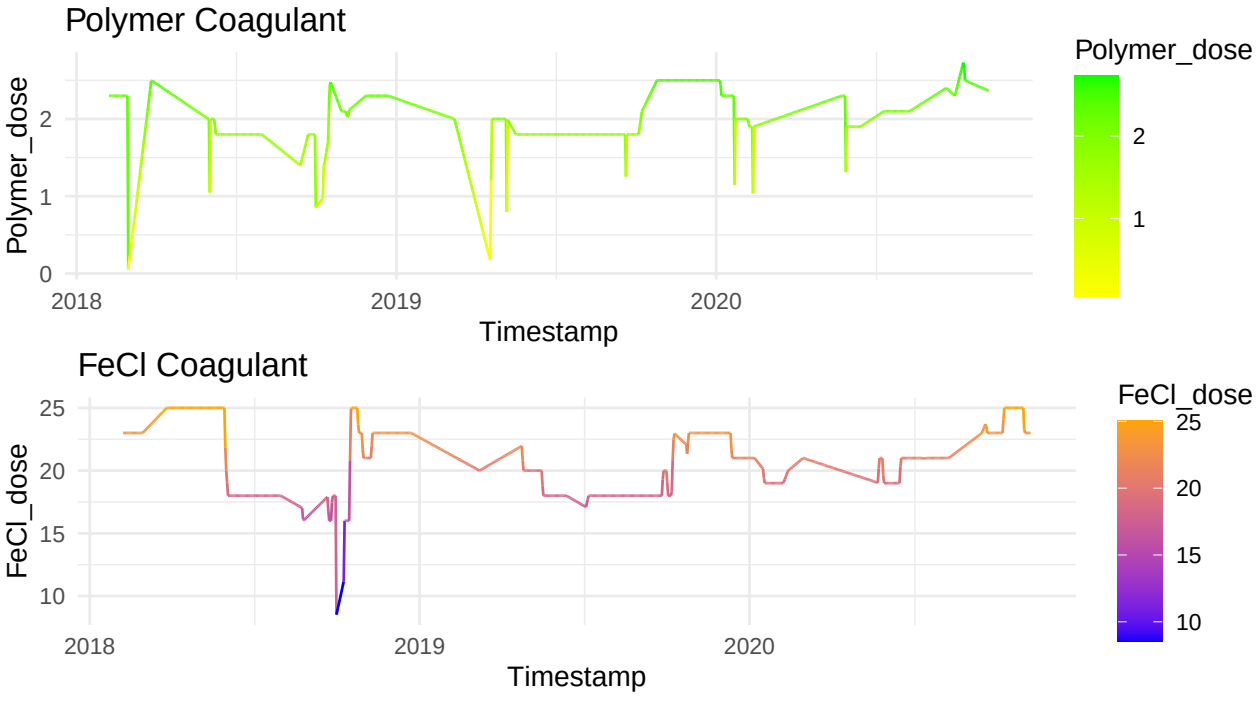
Primary Raw Water Characteristic Data

The main four raw water characteristics used in this analysis are pictured below. Turbidity displays the highest levels of fluctuation and noise, followed by conductivity. pH and temperature tend to display fairly consistent annual cycles without significant noise. According to the treatment operator who provided the data, temperature and pH are important for the treatment process, so less noise could be advantageous for modeling.



Coagulant Dose Data

The following are the dosing over time of Ferric Chloride (primary coagulant) and the polymer coagulant. Note the long periods of no dosing change, some periods of sharp dips and jumps, and periods of steady increase and decrease. This is how operators alter coagulant doses day-to-day. This gives a picture of some of the challenges that may be present in trying to predict this coagulant dosage based on water characteristics.



Predictor Selection

Based on preliminary analyses, we will focus on the following dose predictors, which have the greatest time availability and fewest data gaps out of our available raw water data:

- **pH** (strongest relationship)
- **Temperature** (strongest relationship)
- **Conductivity** (some noise, weaker)
- **Turbidity** (lots of noise, weaker)

Overall, it seems like temperature and pH have the strongest relationships to chemical dosing. This is consistent with feedback from treatment plant operators. It is not immediately clear which relationships could make effective predictors. Turbidity is highly variable and likely does not have a strong relationship.

We will focus on **FeCl Dose** as the outcome variable for our modeling efforts. Polymer dose is not changed significantly in plants, as it has less direct impact on the coagulation performance and serves as a secondary "helper" in particle formation to the ferric chloride.

Added Variables

To capture some of the time-dependence likely present in our dataset, we introduced **lagged variables** for each of our main raw water characteristics. We introduced two lagged variables, representing the water characteristics for the day before and two days before.

We also introduced a **categorical predictor** indicating whether the dose increased, decreased, or stayed the same compared to the day before. This will be used for dose change prediction modeling (classification), which is detailed below.

Modeling

Summary of Modeling Efforts

Modeling Priorities

The main priorities in this modeling effort are *interpretability*, and effective capturing of *nonlinear behavior*. Interpretability is important because to trust a model, a treatment plant operator will want to understand the inner workings of how it is making its predictions. Nonlinear behavior is important because the chemical process involved in the treatment process are highly complex so a simple linear model will likely not capture these chemical relationships effectively. We recognize that these two goals can be at odds, due to the harder-to-interpret nature of highly nonlinear models, so we will attempt to choose models that balance these two objectives.

Training and Testing Datasets

Our total dataset spans 2018-2020, so we will use data from 2018-2019 as the training dataset, and 2020 as the testing dataset. This accomplishes approximately a 70/30 training/testing split for our available data.

We recognize that there may be limitations to this train/test split, since 2020 was both the height of the COVID-19 pandemic, and a year of above-average numbers of forest fires. The former of these could have impacted the funding and staffing of the water treatment plant, along with consumptive use patterns in the regions serviced by the plant. The fires could have directly impacted water quality with more dissolved ash in the water supply.

These could have unforeseen impacts, but it seems upon preliminary assessment that these factors should not impact plant performance significantly. Water treatment is critical infrastructure, so should not be substantially impacted by shutdowns. Additionally, it is already only staffed by a few people at a time. Shifting consumptive use patterns would not impact reservoir water quality.

Reservoirs also tend to be fairly stable water sources, in terms of quality, so the additional ash that may have come from the fires is likely negligible. We observed some sparse datasets available, including dissolved organic carbon and total organic carbon measurements, and 2020 was not significantly different from 2018 or 2019 in this way, which could be a helpful indicator.

With this in mind, this split will be maintained. For future efforts, with more data, different splits will be tested, but with this available data, this seems like the best approach.

Models Used

The following models will be used for the two main tasks at hand. Models that work well with numerical outcomes will be used for coagulant dose prediction. Models that are effective for classification will be used for coagulant dose change prediction, since this is a three-category classification problem (increase, decrease, stay same).

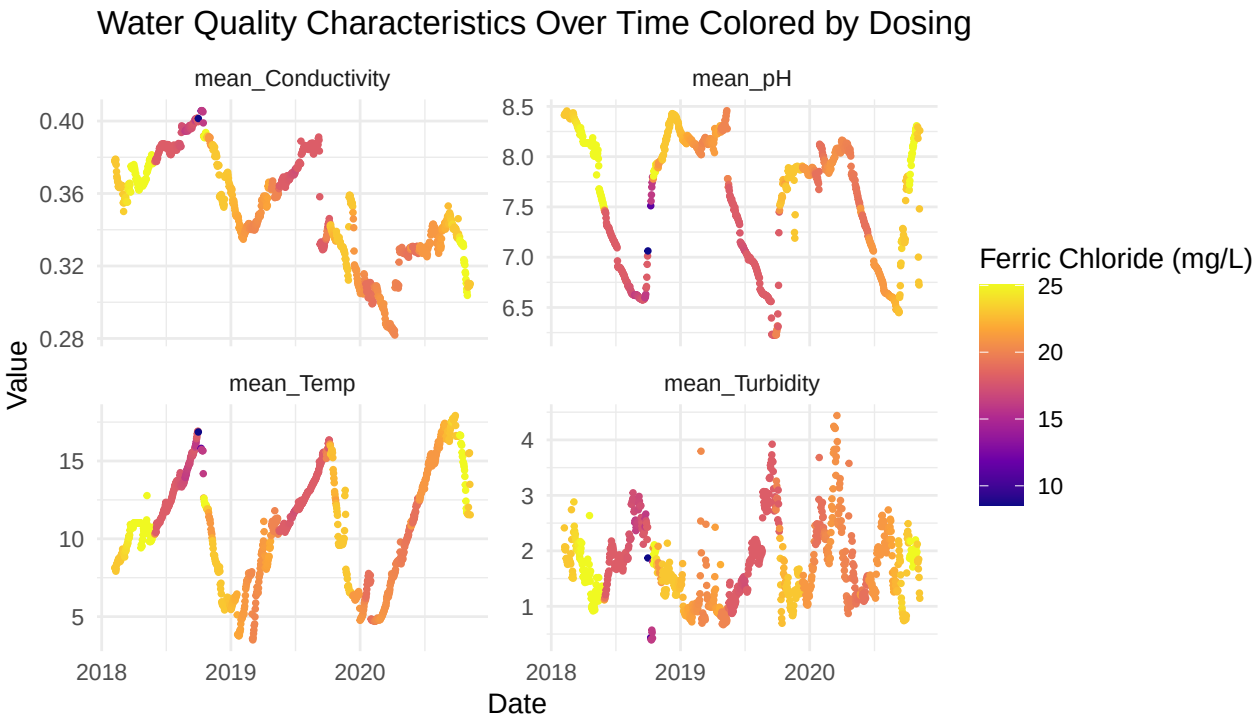
- **Coagulant Dose Prediction**
  - Linear Regression Models
  - Tree-Based Models
  - Generalized Additive Model
- **Coagulant Dose Change Prediction**
  - LDA and QDA Classification
  - Multinomial Logistic Regression
  - Random Forest Classification

Test MAE and RMSE will be used to assess dose prediction models, with best model chosen based on RMSE. Test accuracy will be used to assess dose change prediction models.

Dose Prediction Models

Overview of Dose Prediction

The goal of the dose prediction modeling is to predict FeCl dosage from the raw water characteristics. The following plot helps illustrate the data that will be used for the dose prediction model. This plot does not reveal any clear trends in dosage relative to any of the water treatment variables (dose varying with a water quality metric), indicating this could be a challenging modeling effort.



Linear Regression

To model FeCl dosing using linear regression, we fit six different linear regression models that varied in their choice of predictors and inclusion of differing interaction terms. The models incorporated current and lagged values of temperature, pH, conductivity, and turbidity, allowing us to account for delayed effects in the system from previous days. Several interactions were tested, including temperature-pH and temperature-conductivity, to capture potential nonlinear relationships between water quality variables. Turbidity was ultimately dropped from the final models, as it consistently preformed poorly as a predictor for FeCl dosing.

Displayed are both a summary of the fitted regression models, and an example plot of the actual versus predicted dosing, presented for the best-performing linear model, Model 3. This better illustrates how far the predictions are from the actual dosing.

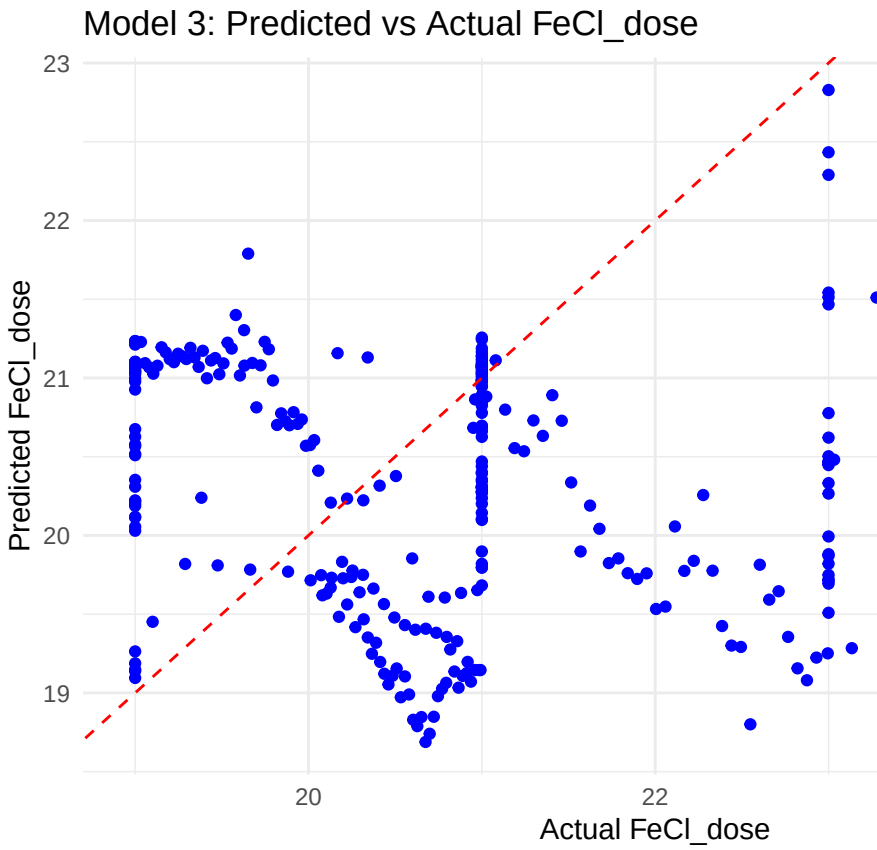
All predictors:

Generate model results

```
##     Model  RMSE   MAE                          Description
## 1 Model 4 1.672 1.421 Cond × pH + Temp × Cond interactions
## 2 Model 3 1.891 1.516              Temp × Cond interaction
## 3 Model 5 2.350 1.941                     Only Temp and pH
## 4 Model 0 2.971 2.580                    Main effects only
## 5 Model 2 3.201 2.784 Temp × pH + Cond × Turb interactions
## 6 Model 1 3.483 2.938                Temp × pH interaction
```

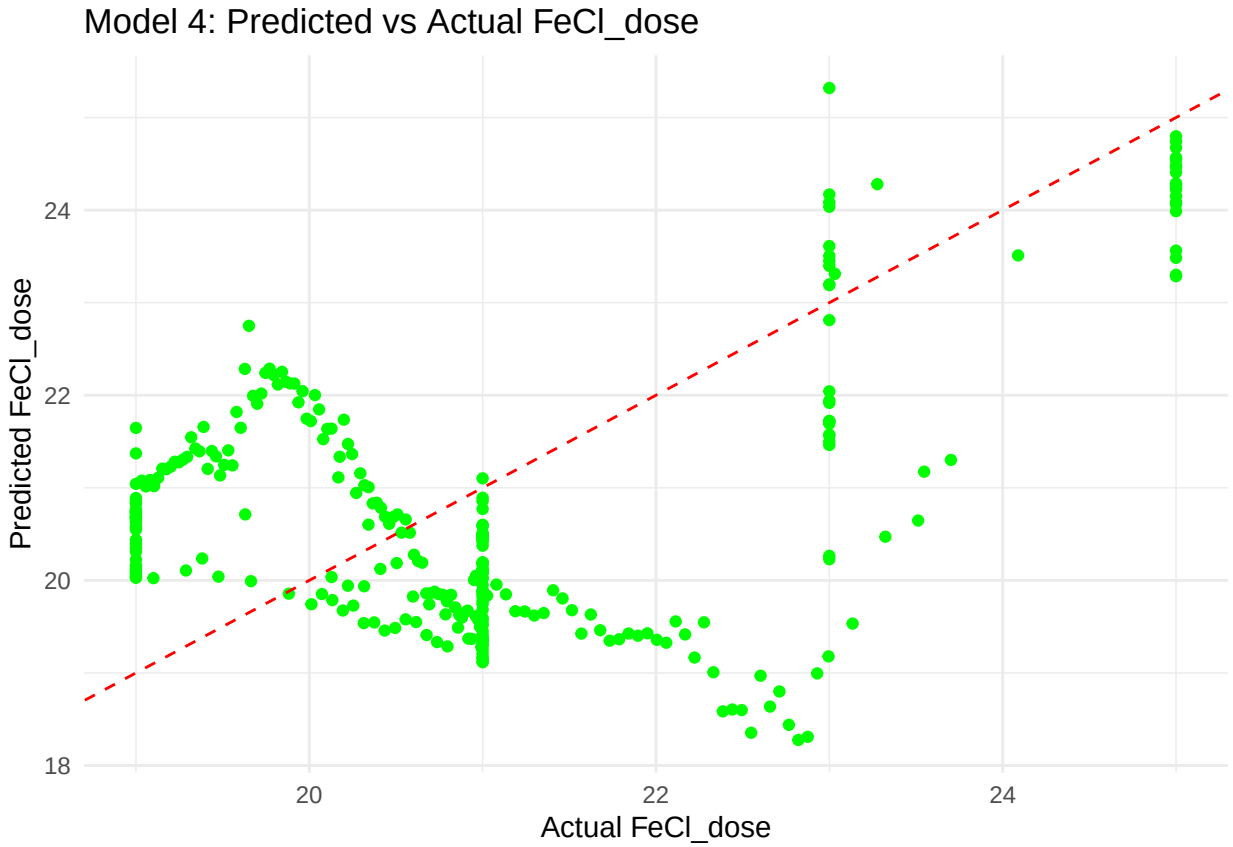
Selecting the best RMSE:

```
## [1] 1.672
```



Display Plots of models' performance on test data:

Model	RMSE	MAE	Description
Model 4	1.672	1.421	Cond × pH + Temp × Cond interactions
Model 3	1.891	1.516	Temp × Cond interaction
Model 5	2.350	1.941	Only Temp and pH
Model 0	2.971	2.580	Main effects only
Model 2	3.201	2.784	Temp × pH + Cond × Turb interactions
Model 1	3.483	2.938	Temp × pH interaction



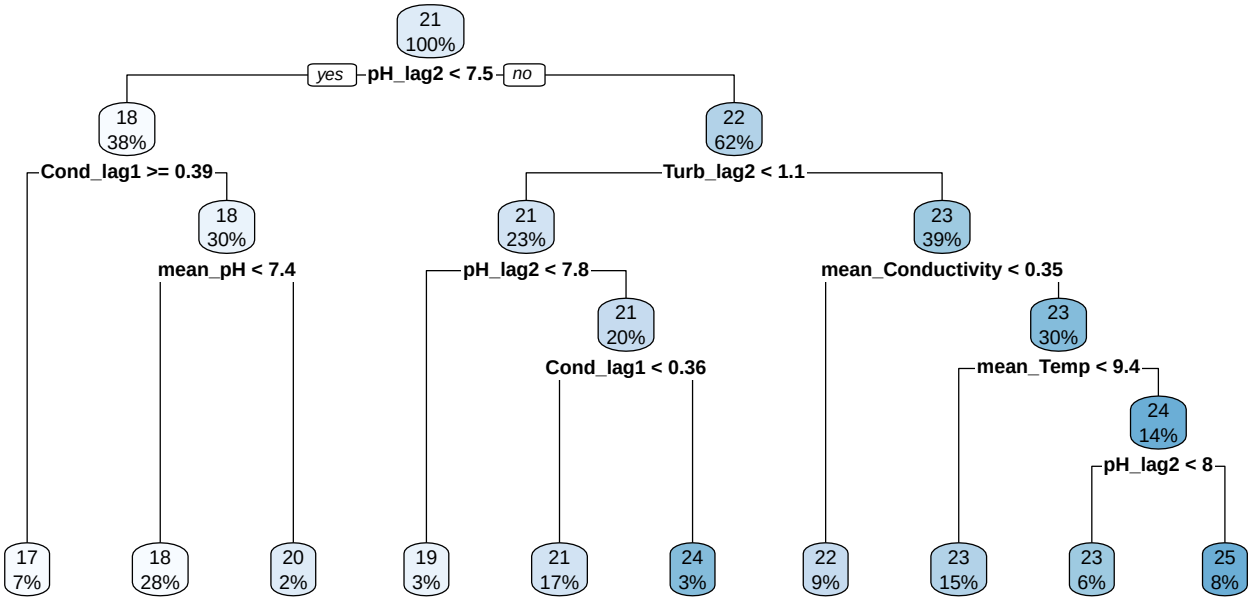
Model performance was evaluated using RMSE and MAE on a 2020 test set, and varied substantially depending on the selected predictors and interactions. There is evidence of other factors influencing the dosing as many of the predictors in our data set proved to be not be as influential as we originally had suspected. This is reflected through the model performance in which the model with temperature-pH and temperature-conductivity being the best, but its predictions still deviated considerably from the actual FeCl dosing values. As seen in the plots and table above, none of the models performed particularly well, and there was a ton of variance within each model and how well it predicted.

Decision Tree

The benefit of a tree-based model in this application is that it is easily interpretable. It could serve as a decision-making tool for a treatment plant operator:

- The operator sees a change in raw water conditions
- Obtains a recommended dose based on the decision tree
- Easily interprets the "why" of the dosing recommendation by following the logic of the decision tree
- Understands which variables were most important in deciding on a dose

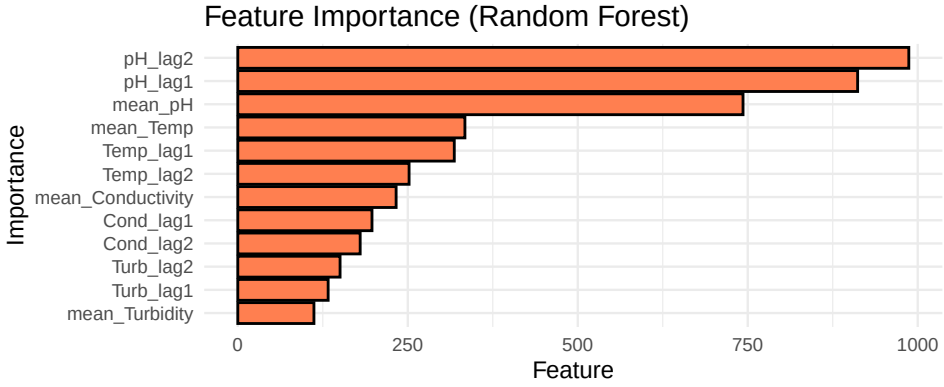
A decision tree was fit with a depth of 5, and 50 observations required to split. These parameters were varied but it did not significantly improve prediction. The model split first based on pH, which was mentioned by treatment plant operators as an important predictor. However, the next splits were on lagged conductivity and turbidity measurements. This was surprising, and could likely be due to the noise present in the data. (Possible result of overfitting.) This model performed more poorly than the linear models. See below for model performance metrics and a visualization of the final decision tree.



Metric	Value
RMSE	2.756
R-Squared	0.012
MAE	2.484

Random Forest

Random forest was used and was the next logical step to try to better capture any nonlinear patterns in our data. Multiple different tree amounts were attempted, but ultimately they all performed similarly, and worse than the linear models. Although the random forest was unable to model the data, we were able to determine which features had the most importance. A graph was created to visualize the feature importance for this dataset. Notably, pH and the pH lag variables were the most important. Temperature and the temperature lag variables followed, which is similar to what we concluded from other models and our exploratory analysis.



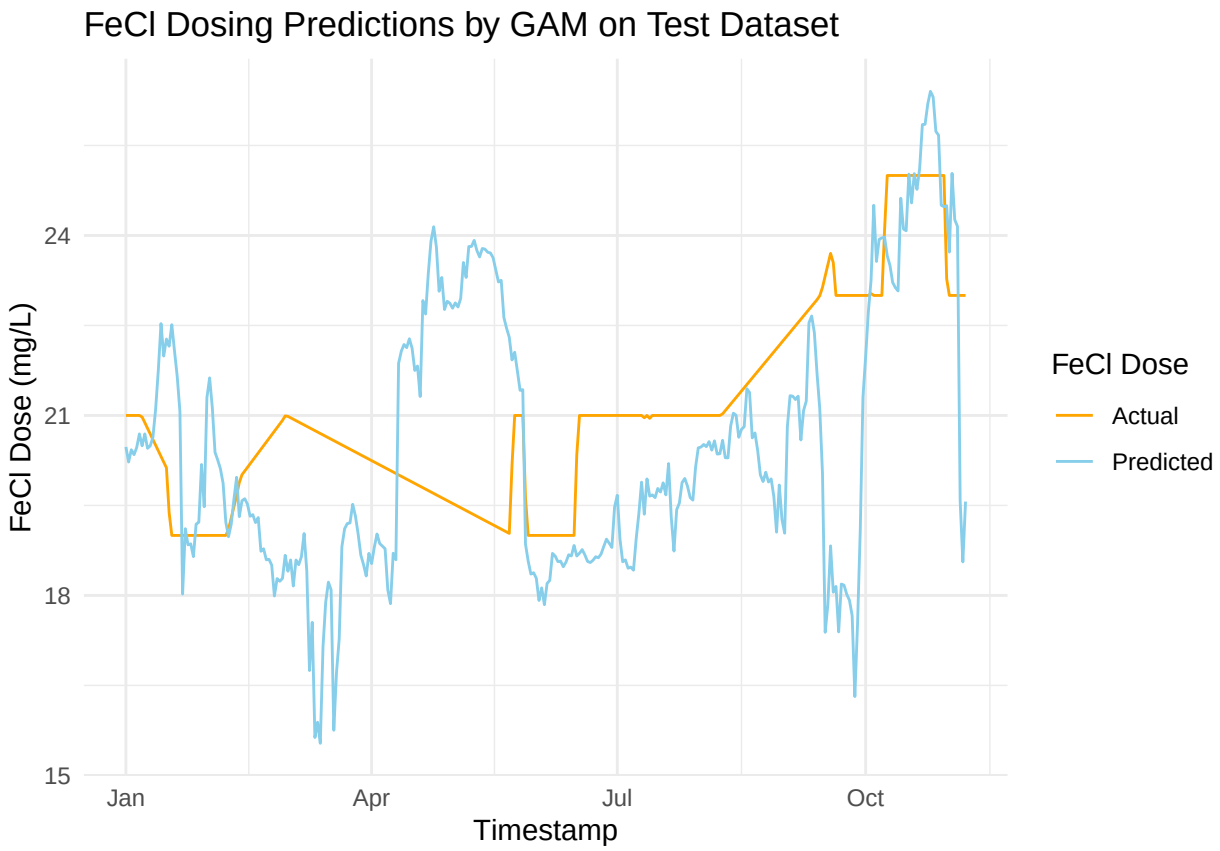
Metric	Value
RMSE	2.811
R-Squared	0.021
MAE	2.283

Generalized Additive Model (GAM)

Next, a GAM was fitted to the modeled data. This model uses splines to model nonlinear relationships between predictor and response. The hope would be that this flexible model form could help capture the nonlinearity present in these chemical processes, for effective prediction. See below for a plot of how the GAM performed on the test dataset.

```
## # A tibble: 13 x 5
##    term                    edf ref.df statistic    p.value
##    <chr>                 <dbl>  <dbl>     <dbl>      <dbl>
##  1 s(Polymer_dose)        8.64   8.95  41.9      0
##  2 s(mean_Temp)           8.36   8.81   5.50     0.0000596
##  3 s(mean_pH)             8.48   8.86   3.53     0.000331
##  4 s(mean_Conductivity)   5.00   6.18   2.59     0.0160
##  5 s(mean_Turbidity)      8.10   8.76   5.27     0.00000134
##  6 s(Temp_lag1)           1.94   2.57   0.702    0.634
##  7 s(Temp_lag2)           7.79   8.57   3.25     0.0132
##  8 s(pH_lag1)             1.00   1.00   0.00167  0.967
##  9 s(pH_lag2)             8.14   8.74   3.59     0.000296
## 10 s(Cond_lag1)           1.00   1.00   0.268    0.605
## 11 s(Cond_lag2)           7.53   8.34   4.27     0.000235
## 12 s(Turb_lag1)           1.78   2.28   0.511    0.553
## 13 s(Turb_lag2)           5.31   6.57   2.85     0.00701
```

```
## # A tibble: 1 x 9
##      df logLik   AIC   BIC deviance df.residual  nobs adj.r.squared  npar
##   <dbl>  <dbl> <dbl> <dbl>    <dbl>       <dbl> <int>         <dbl> <int>
## 1  74.1  -800. 1750. 2089.     417.        608.   682         0.908   118
```

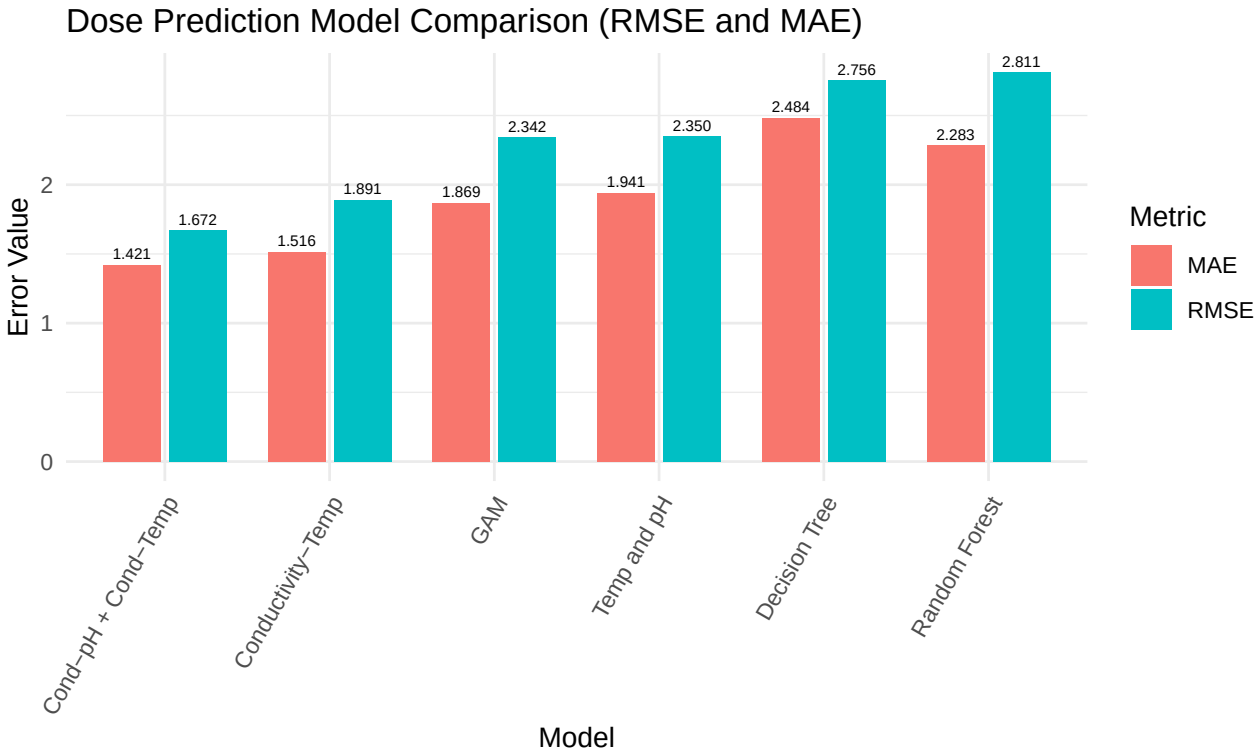


Overall, the GAM was able to vaguely follow trends in dosing for the testing dataset, but did not perform overly well, with MAE and RMSE of 1.8 mg/L and 2.3 mg/L, respectively. The model flexibility did not turn out to yield the hoped-for results, though it performed similarly to the standard linear models. It did out-perform both of the tree-based models.

There does seem to be a faint trend between the model performance and the observations in a postive slope direction. This model may perform better if we had more years at our disposal to train the GAM on, but how much better is highly speculative because of the variation of predictions from the model in contrast to the observations. More data could be what this model needs.

Summary of Dose Prediction Models

Overall, none of the dose prediction models performed particularly well or were able to follow the changes in chemical dose present in the modeling. The top three models by RMSE were the models involving interactions between conductivity and temperature/pH, and the GAM. However, the best-performing models still differed from the actual dosing by an average of at least 1.5 mg/L based on MAE, which is a significant dosing deviation that could hinder proper functioning of a treatment plant if implemented.



Next, we will move on to dose change prediction, and see how its performance compares to this model estimation of exact doses.

Dose Change Prediction Models

Overview of Dose Change Prediction

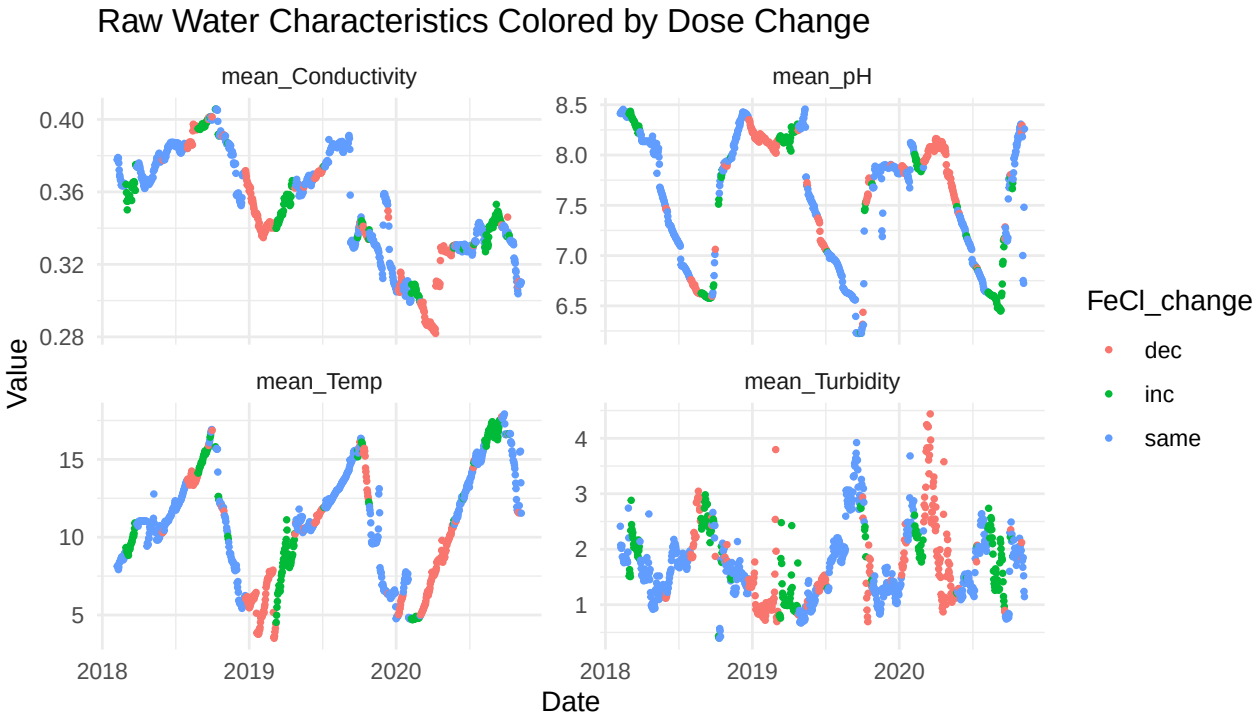
The goal of dose change prediction models is to predict whether coagulant dose increased, decreased, or stayed the same based on raw water characteristics.

As mentioned previously, we added two variables called `FeCl_change` and `Polymer_change` that indicate the chemical dose's direction of change.

This will allow classification methods to be applied. Additionally, if effective, this could provide a more helpful tool for operators. Instead of attempting to predict a specific dose, it could give a general direction

for an operator to attempt a dose change. This is how plants work in practice: an operator uses some tool (a water test, a reading from a device, etc.), and uses that to make a dose change. They then check important treatment metrics and make any needed further adjustments.

The following is a helpful visualization of how dose change varies over time with each of the key raw water characteristics. As with dose prediction, there are not clear patterns of when dose is increasing, decreasing, or staying the same relative to water quality metrics. This will most likely make the data difficult to separate via classification methods.



For each of the following classification models, different combinations of predictor variable sets were experimented with. It was found that pH and temperature, with their associated lag variables, produced the highest test accuracies. Turbidity and conductivity did not improve test accuracy, so were not included in the final model that is represented by the confusion matrices and accuracies below.

Linear Discriminant Analysis (LDA)

LDA uses linear combinations of features to perform dimensionality reduction and create better separation between classes of data. Thus, we hope that for this hard-to-separate data, the dimension reduction could assist, though there are not many dimensions to start with, so we could foresee it still being highly limited.

TEST ACCURACY: 43.7%

Table 3: LDA Confusion Matrix

True	dec	inc	same
dec	26	2	74
inc	22	6	44
same	17	16	104

Quadratic Discriminant Analysis (QDA)

QDA is similar to LDA but incorporates quadratic terms, not just linear terms. This could potentially help address the nonlinearity in our dataset.

However, based on test accuracy and the confusion matrix, the QDA model performed just as poorly as the LDA, and slightly worse. See results below.

TEST ACCURACY: 35.7%

Table 4: QDA Confusion Matrix

True	dec	inc	same
dec	71	2	29
inc	57	5	10
same	86	16	35

Multinomial Logistic Regression

Multinomial logistic regression is an extension of logistic regression that allows for multi-category classification, by setting one category as a "reference category" and calculating probabilities of other categories relative to that category. Overall, this model tended to predict the best on the test dataset. Looking at the confusion matrix, it tended to lean towards predicting "same" when it made an error, which is more acceptable than choosing the opposite direction of dose change, though it made that error as well. However, it is still not an exceptional model by any means. See below for confusion matrix and accuracy.

TEST ACCURACY: 46.0%

Table 5: Multinomial Regression Confusion Matrix

True	dec	inc	same
dec	26	0	76
inc	22	3	47
same	16	7	114

Random Forest

Finally, a Random Forest model was fit to this data. The standard number $\sqrt{P}$ predictors were sampled at each split, and 500 trees were grown. The number of trees was varied but it did not change prediction accuracy.

This model performed second-worst out of the classification models. We believe that this is mainly due to overfitting. Random forest models tend to overfit on noise-heavy data, so this model performed perfectly on the training dataset and very poorly on the test dataset. See accuracy and confusion matrix, below.

TEST ACCURACY: 39.2%

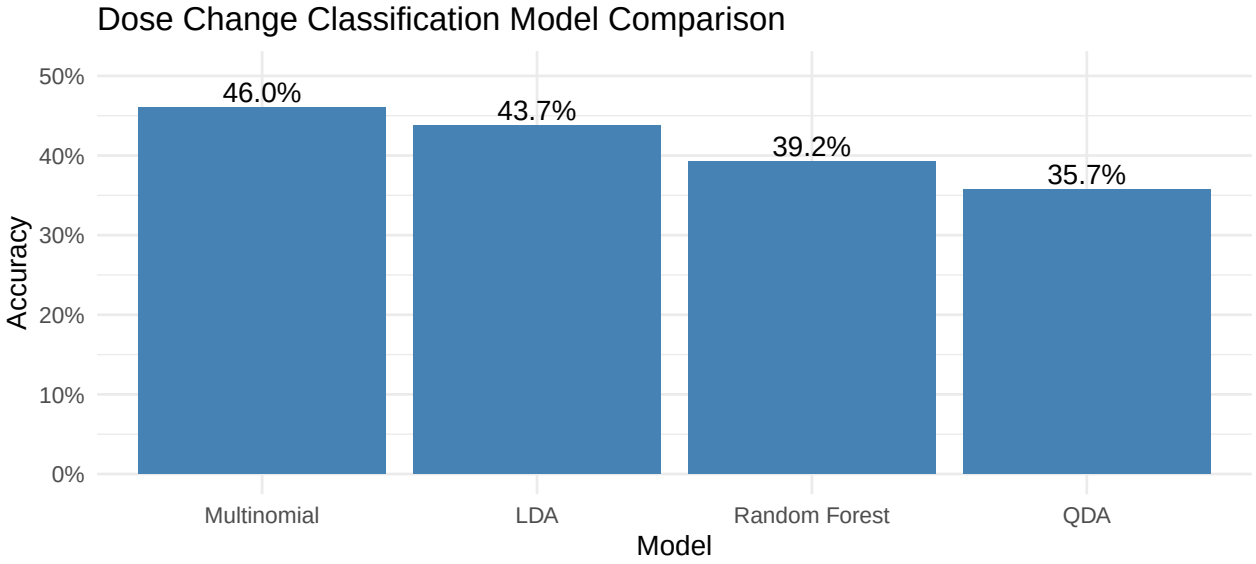
Table 6: Random Forest Confusion Matrix

True	dec	inc	same
dec	37	6	59
inc	17	7	48
same	37	22	78

Summary of Dose Change Prediction Models

These models were ineffective at predicting the direction of dose change. This aligns with initial expectations, since the data was shown to be extremely scattered and difficult to separate in the time series plots shown at the start of this section. All models performed very poorly, so it is hard to compare which models may have

performed "best." At the surface, multinomial performed best, likely due to it using linear combinations of predictors (which aligns with the most effective dose prediction methods). Random forest was the most prone to overfitting and worst-performing, which aligns with its tendency to overfit on noise-heavy data. Model accuracies are summarized below.



Conclusion

Overall, predicting chemical doses based on these raw water characteristics proved to be ineffective.

The dose prediction models produced high RMSE's, either being unable to capture the complexity of the dosing, or overfitting to the noise-filled signals of the input data. The dose change prediction models, as predicted by observation of the timeseries plots, yielded extremely low accuracies, due to this being a dataset not separable by any one of these four input water quality characteristics.

Some of the most likely reasons for this are as follows:

- **Limited dataset size**: our dataset was limited in time, forcing us to use a small training and testing dataset that limited our ability to choose other years for testing. This likely produced errors specific to the year that we chose for testing, which could be averted with more data.

- **Limited predictor set**: These four water quality variables were not effective for predicting chemical doses. However, there are other water quality parameters, and other treatment plant performance characteristics such as filter performance, that are used to make dose decisions. Access to these data could significantly improve model performance.

- **Operator-dependence**: Whether doses even change at all is largely dependent on operators. This plant is one of the more advanced in the nation, but some plants leave their dose the same for years at a time. So, our model is in part simulating human behavior, which is much more difficult than a basic model can capture.

- **Complexity of chemical relationships**: The drinking water treatment process is an enormous chemical reaction with a large number of interconnected processes and chemical interactions, so our models likely failed in part to not being able to fully capture these relationships.

- **Other factors**: Other miscellaneous factors can impact water quality, such as reservoir turnover, which is when a reservoir's water level drops, and water quality changes drastically over the course of a few days at a specific time of year. This is hard to predict, and our model could not capture this.

Ultimately, water treatment is both a science and an art, and operators make decisions based on variety of inputs: chemical waste production, filter performance, or lab tests, to name a few. Future modeling efforts should incorporate more varieties and time ranges of data to attempt to capture some of these complexities.