Prediction of Drinking Water Treatment Process Chemical Doses

Dawson Carney, Reece Carmody, Ethan Schilling, Joshua Tobey

17 December 2025

Contents

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

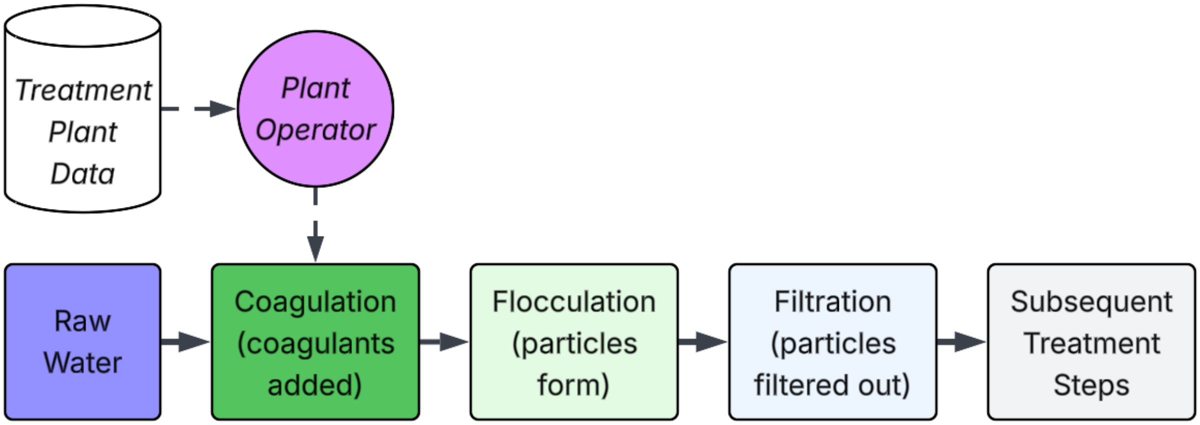
This project will focus on Step 1. These chemical doses are one of the most important features of the treatment process that an operator can change to improve performance.

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 Data
Plant Operator
Raw Water
Coagulation (coagulants added)
Flocculation (particles form)
Filtration (particles filtered out)
Subsequent Treatment Steps

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.

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 (of the water)
  - Turbidity
  - Suspended Grain Size Information
  - Alkalinity
  - Hardness
- 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 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.

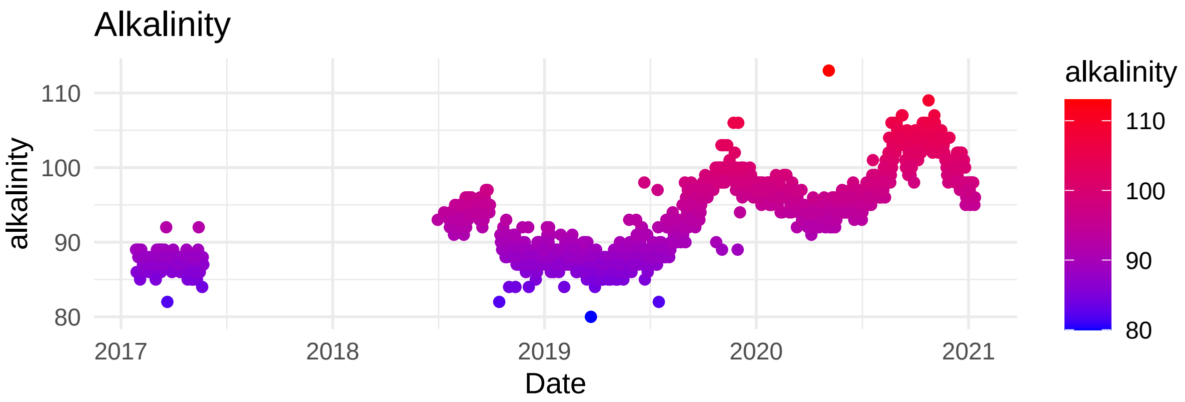
Final Combination: We merged the raw water data (excluding alkalinity/hardness) with the dosing data. We only lost about 10 records in this process, where there was not available dosing data for the raw water measurements.

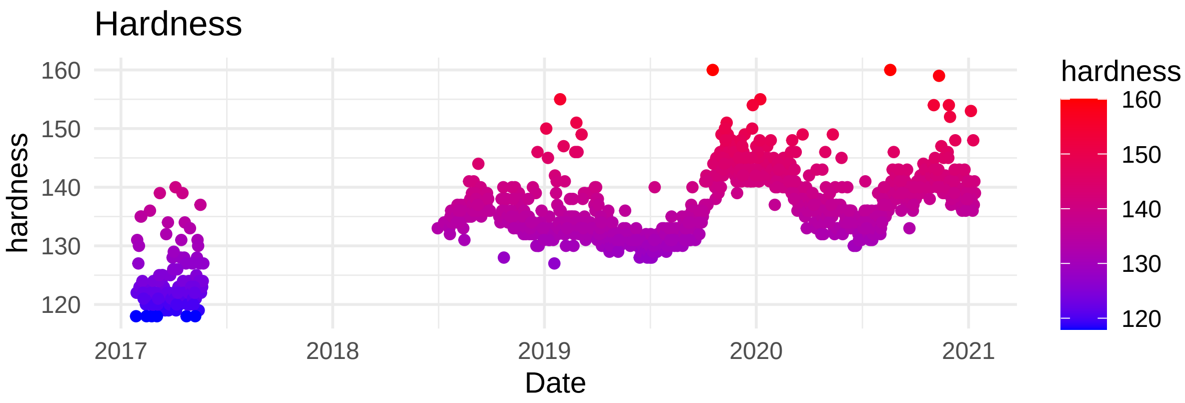
Exploratory Analysis

Some of this cleaned data is presented below.

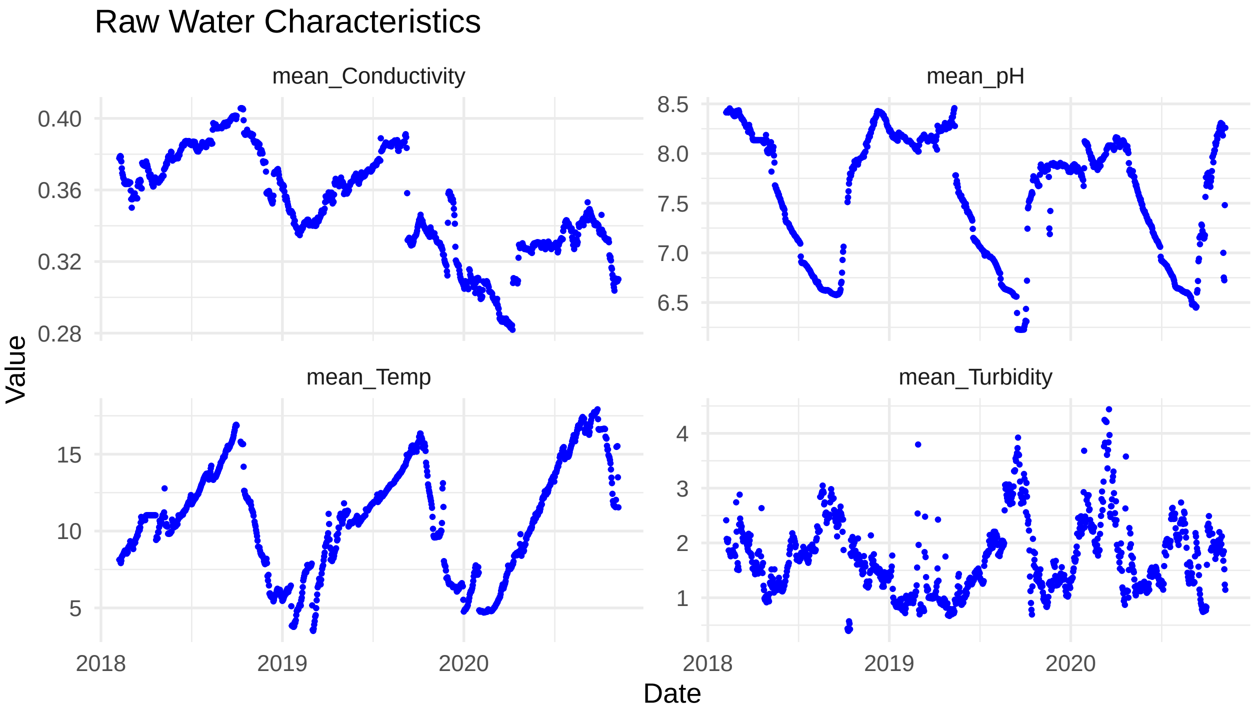
Alkalinity and Hardness Data

Alkalinity and hardness tend to stay fairly steady over time.

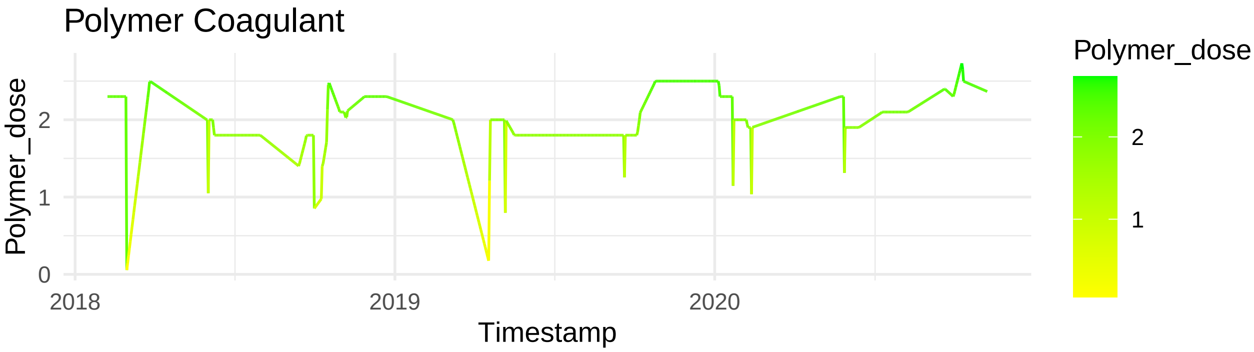


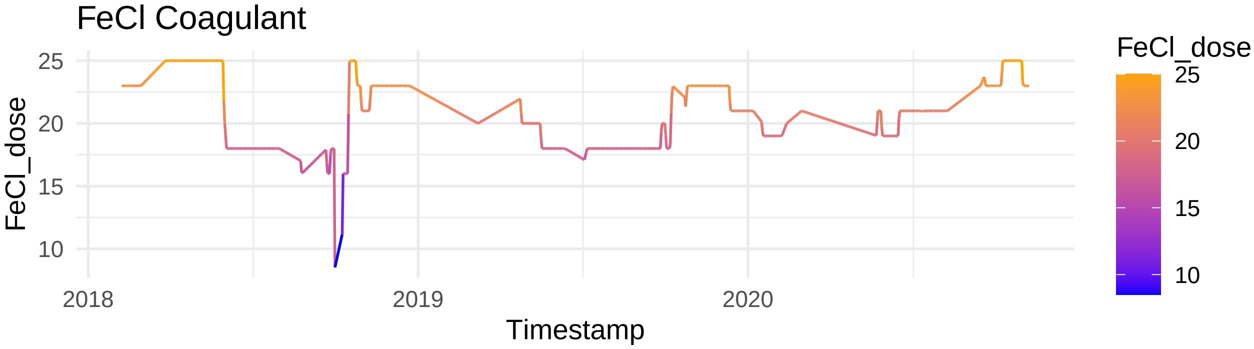


Primary Raw Water Characteristic Data



Coagulant Dose Data





Predictor Selection

Based on preliminary analyses, we will focus on the following dose predictors:

- **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.

Outcome: **FeCl Dose**. (Polymer dose is not changed significantly in plants.)

Added Variables

We introduced **lagged variables** like in the Stock dataset from the homework.

To capture time-dependence of this dataset, we will introduce lagged variables. For each of the main water parameters, we will introduce two lagged variables, giving 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.

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

- Training and Testing
  - Training 2018-2019, Testing 2020
  - Approximately 70/30 split

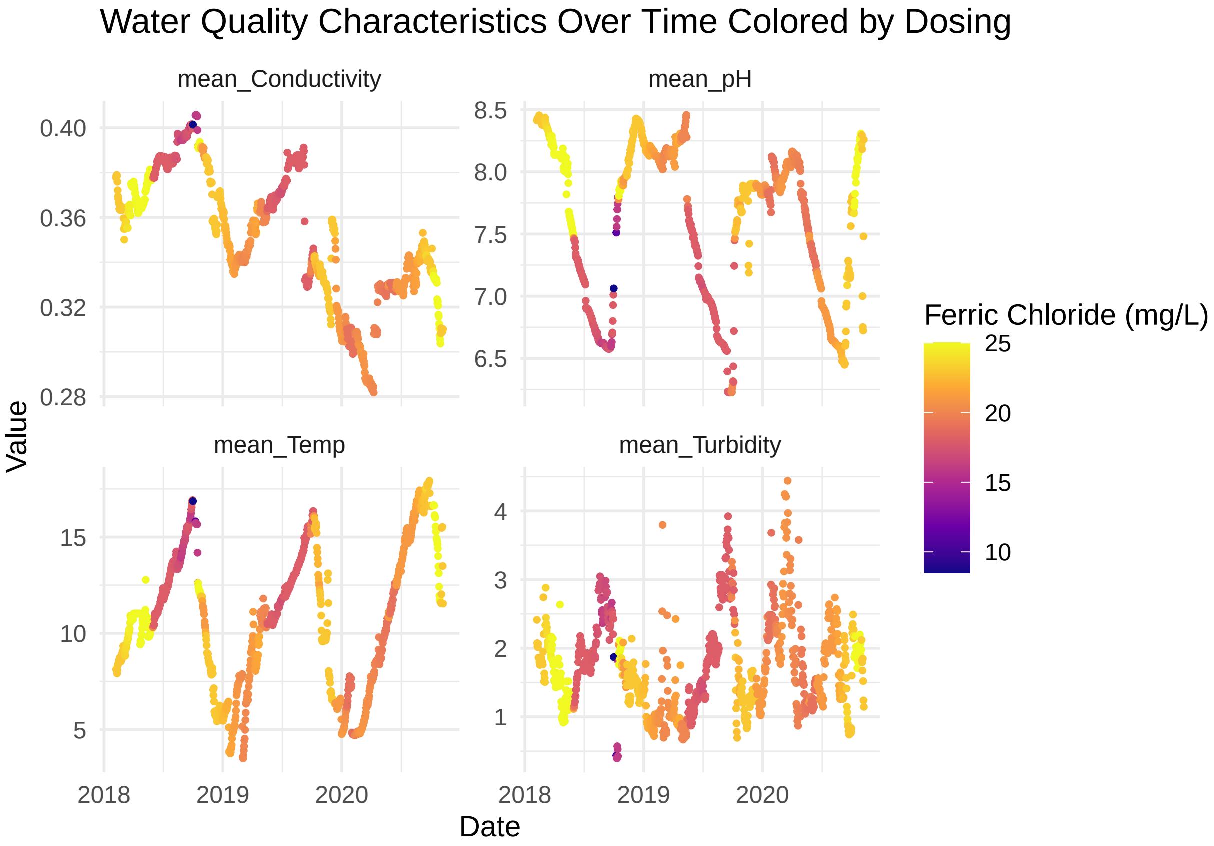
Models Used

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.

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

Dose Prediction Models

Overview of Dose Prediction

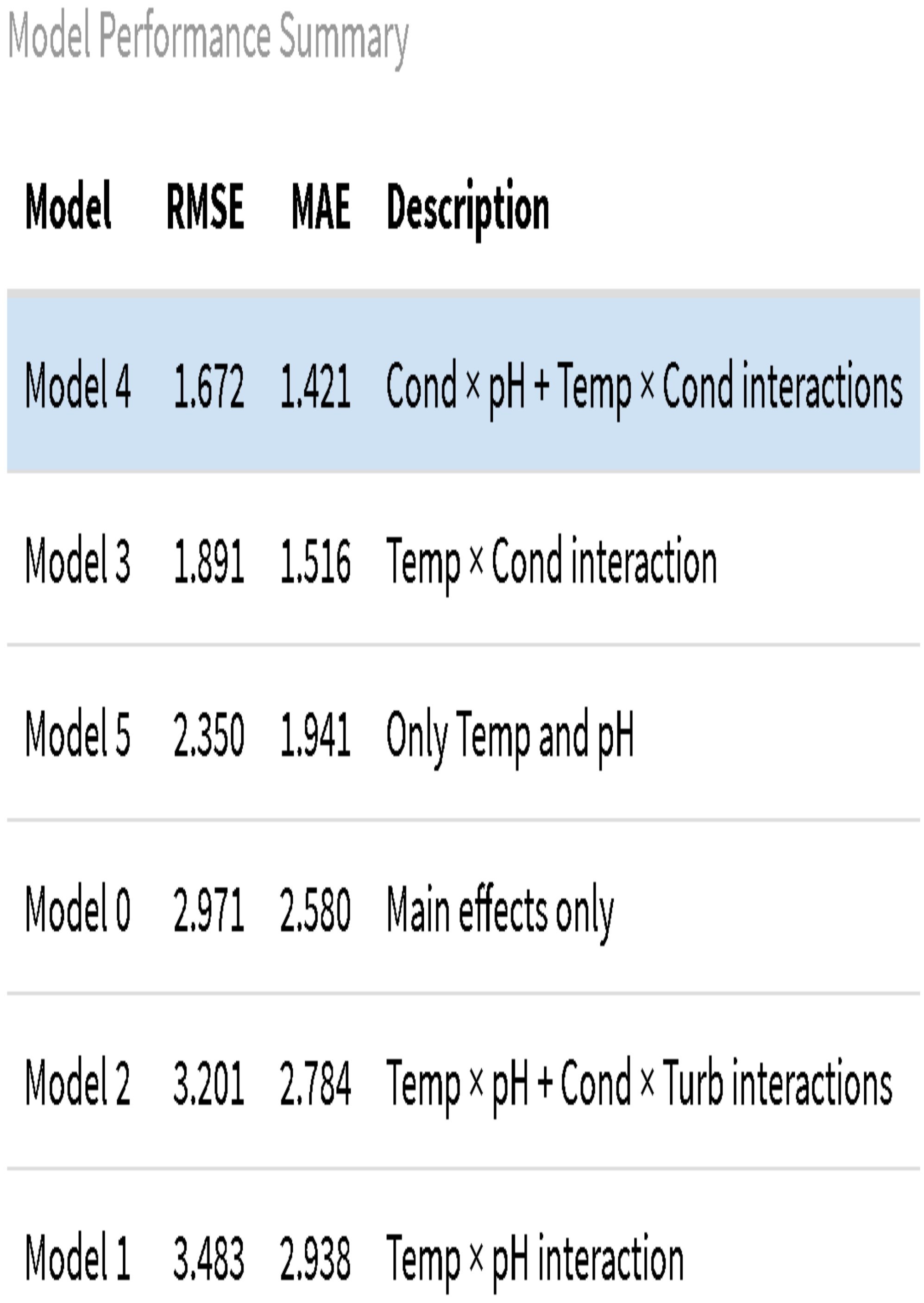


Linear Regression

Ethan's section to fill in

from presentation:

- Created 6 models with different predictors and combinations of interaction terms
- Used lag variables for temp, pH, conductivity, turbidity
- Dropped Turbidity as it was not a good predictor for FeCL dosing
- Difficult to determine what factors of the dataset are driving the predictions
- RMSE and MAE vary widely depending which predictors are selected and the interactions



Model Performance Summary

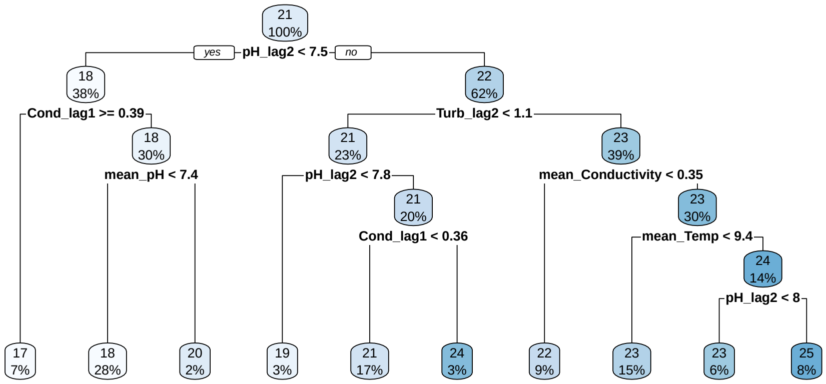
Model	RMSE	MAE	Description
Model 4	1.672	1.421	Cond × pH + Temp × Cond interactions
Model 3	1.891	1.516	Temp × Cond interaction
Model 5	2.350	1.941	Only Temp and pH
Model 0	2.971	2.580	Main effects only
Model 2	3.201	2.784	Temp × pH + Cond × Turb interactions
Model 1	3.483	2.938	Temp × pH interaction

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

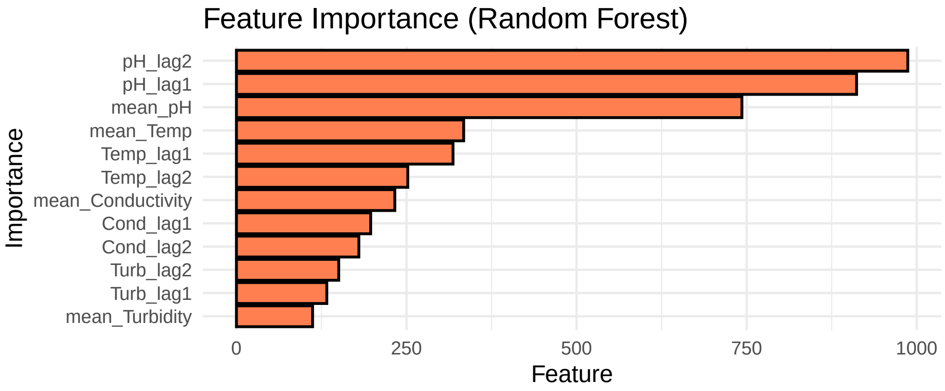
- Obtains a recommended dose based on the decision tree
- Easily interprets the “why” of the dosing recommendation by following the logic of the decision tree
- Understands which variables were most important in deciding on a dose



Metric	Value
RMSE	2.756
R-Squared	0.012
MAE	2.484

Random Forest

(talk about why random forest was used, parameters and predictors used, process of building, etc.)

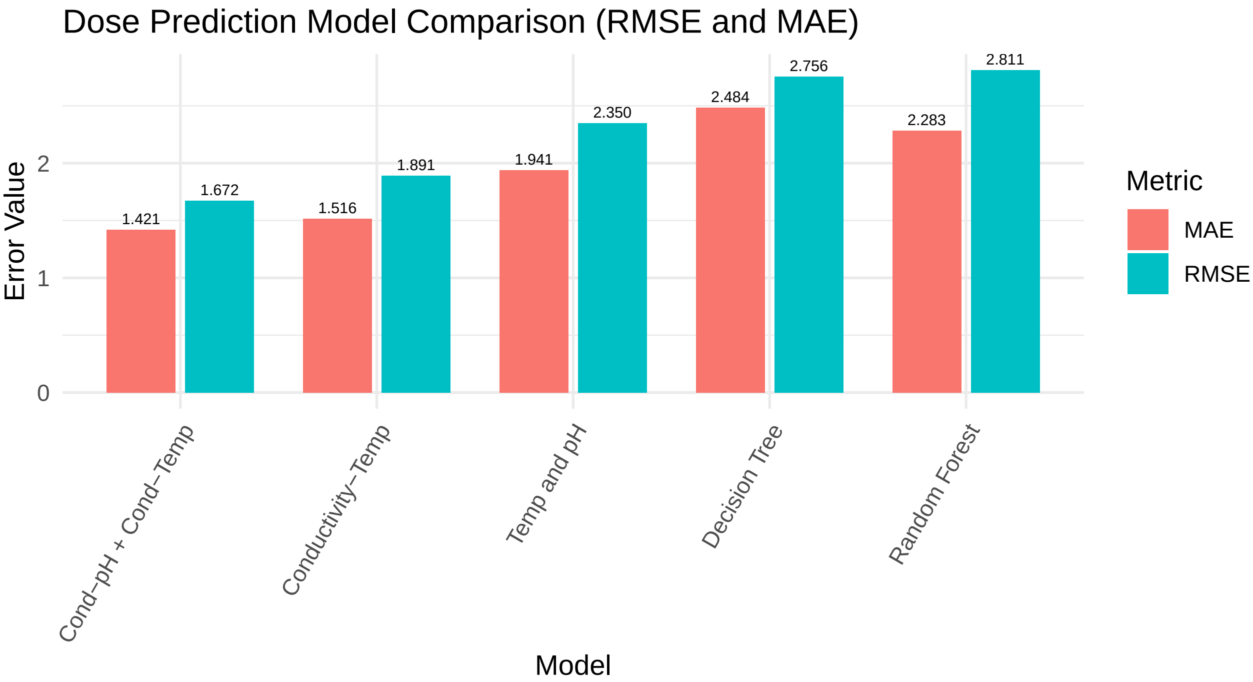


Metric	Value
RMSE	2.811
R-Squared	0.021
MAE	2.283

Generalized Additive Model (GAM)

Josh’s section to insert his GAM

Summary of Dose Prediction Models



Dose Change Prediction Models

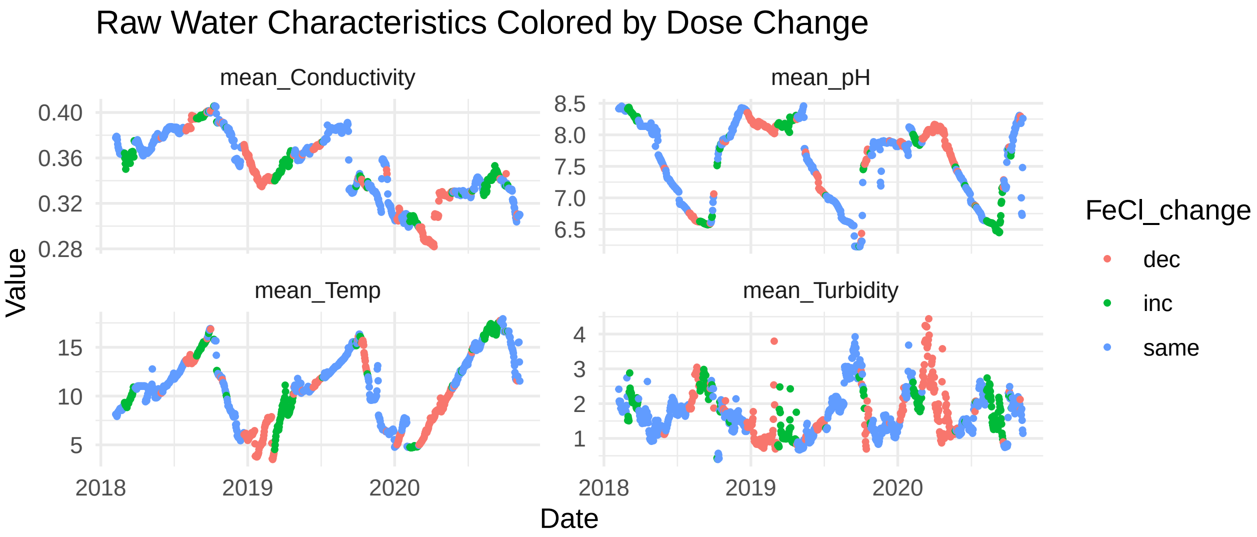
Overview of Dose Change Prediction

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.



Based on this plot, [in essence, we're screwed, but say it diplomatically]

LDA

- Experimented with predictor sets and found that pH and temperature (plus the associated lagged variables) were the main effective predictors.
- Turbidity and conductivity did not improve prediction.

LDA ACCURACY: 43.7%

Table 3: LDA Confusion Matrix

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

QDA

- Experimented with predictor sets and found that pH and temperature (plus the associated lagged variables) were the main effective predictors.
- Turbidity and conductivity did not improve prediction.

QDA 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 Classification

- Experimented with predictor sets and found that pH and temperature (plus the associated lagged variables) were the main effective predictors.
- Turbidity and conductivity did not improve prediction.

MULTINOMIAL ACCURACY: 46.3%

Table 5: Multinomial Regression Confusion Matrix

True	dec	inc	same
dec	31	2	69
inc	22	2	48
same	21	5	111

Random Forest

- Experimented with predictor sets and found that pH and temperature (plus the associated lagged variables) were the main effective predictors.
- Turbidity and conductivity did not improve prediction.

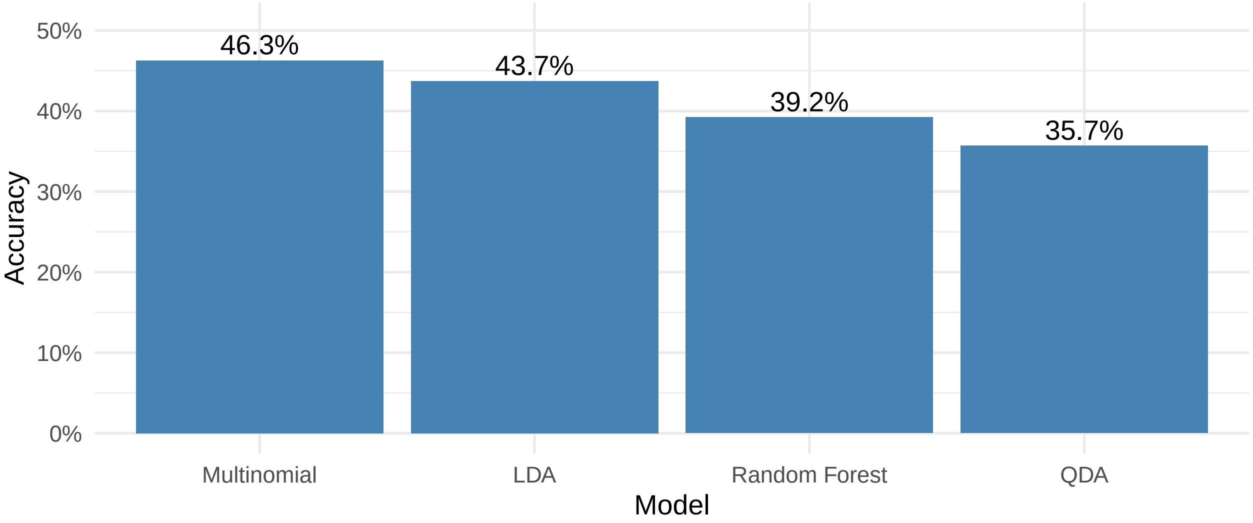
RANDOM FOREST ACCURACY: 39.2%

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

Overall ineffective, and prone to overfitting (especially random forest).
Model predicting no dose change would be more accurate.



Overview of Model Results

(talk about how the models compared, more just synthesizing than making commentary. save commentary for conclusion)

Conclusion

- Predicting chemical doses based on these raw water characteristics proved to be ineffective
- Most likely reasons:
  - Limited size of dataset
  - Limited predictor set
  - Operator-dependent
  - Complexity of chemical relationships
  - Other factors (example: reservoir turnover)
- Operators make decisions based on variety of other factors: chemical waste production, filter performance, lab tests, etc.
- Incorporate more varieties and time ranges of data in future modeling efforts