

Predictive Modeling of Drinking Water Treatment Process Chemical Doses

Dawson Carney, Reece Carmody, Ethan Schilling, Joshua Tobey

5 December 2025

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.

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.

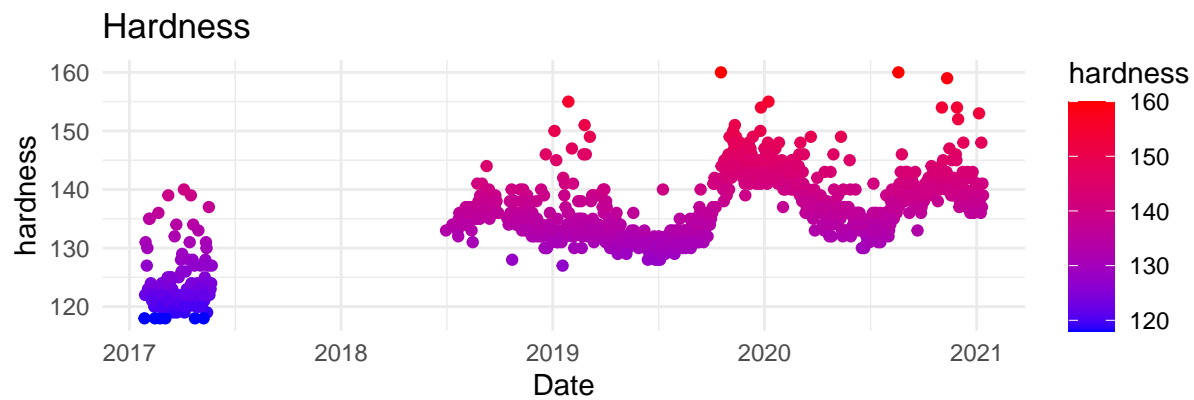
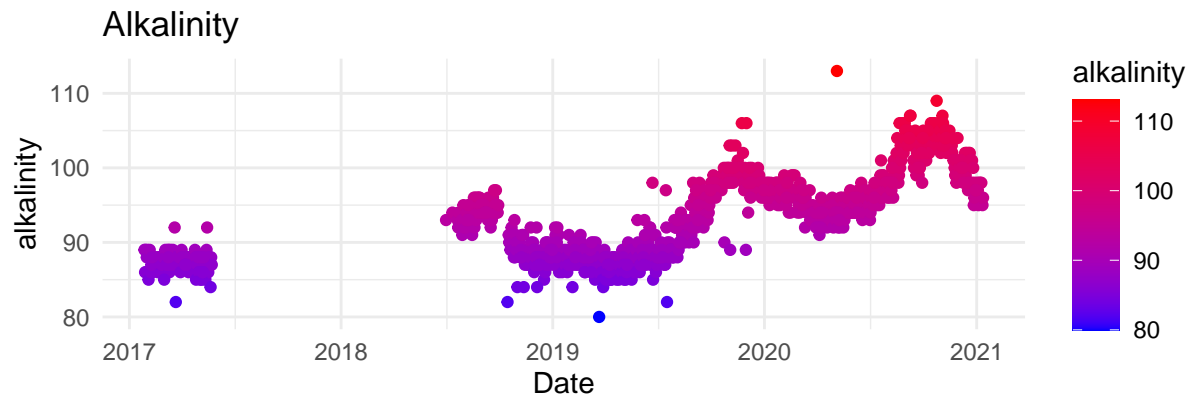
Final Combination: We merged the raw water data (excluding alkalinity/hardness) with the dosing data. We only lost about 10 records in this process, where there was not available dosing data for the raw water measurements.

Exploratory Analysis

Some of this cleaned data is presented below.

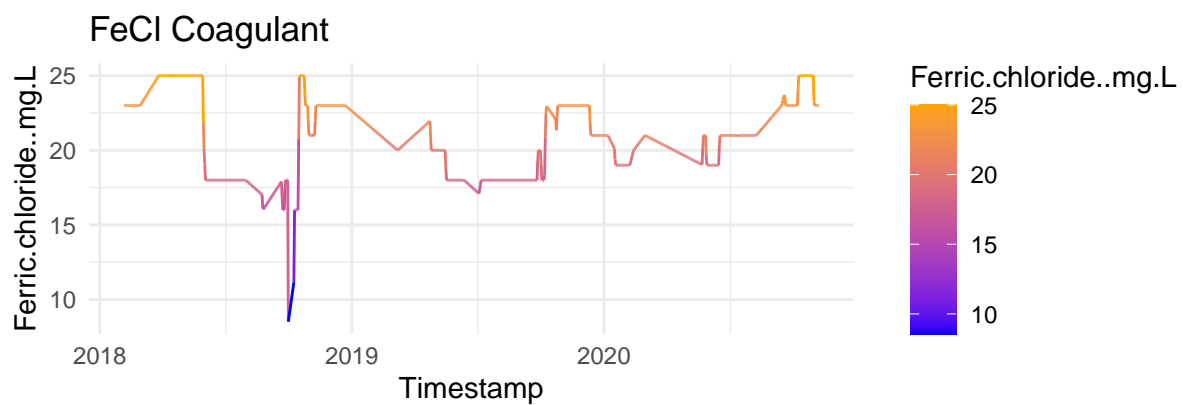
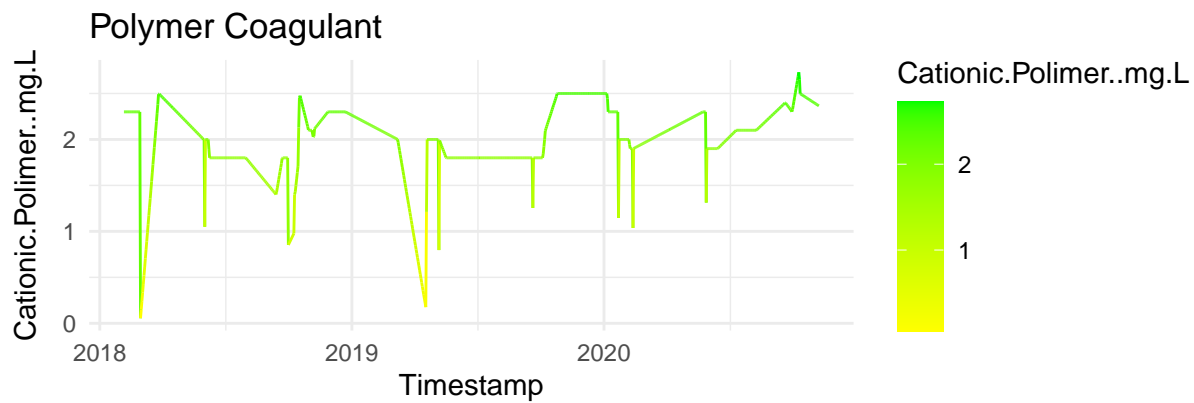
Plots of Key Treatment Plant Variables

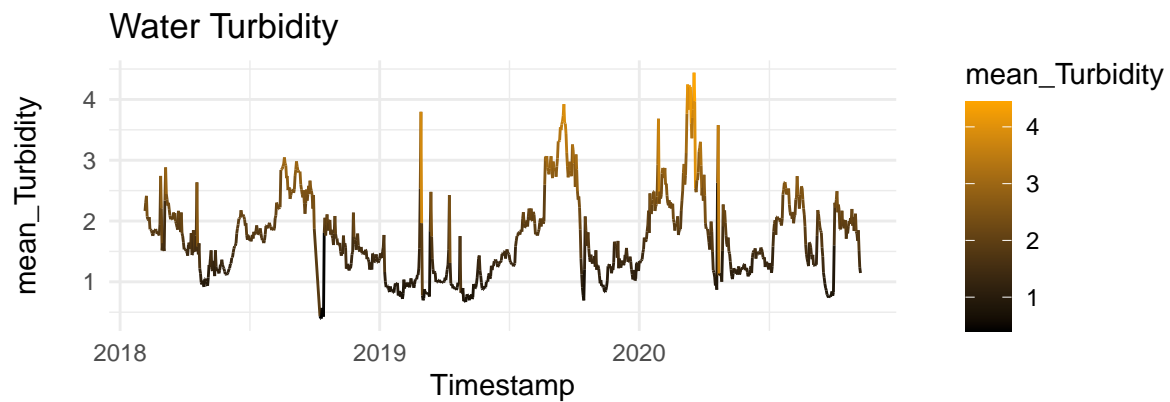
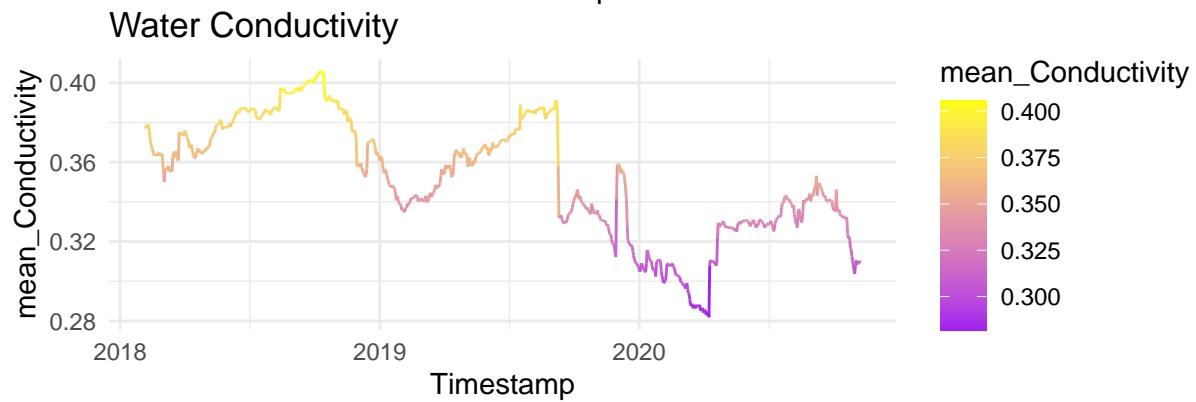
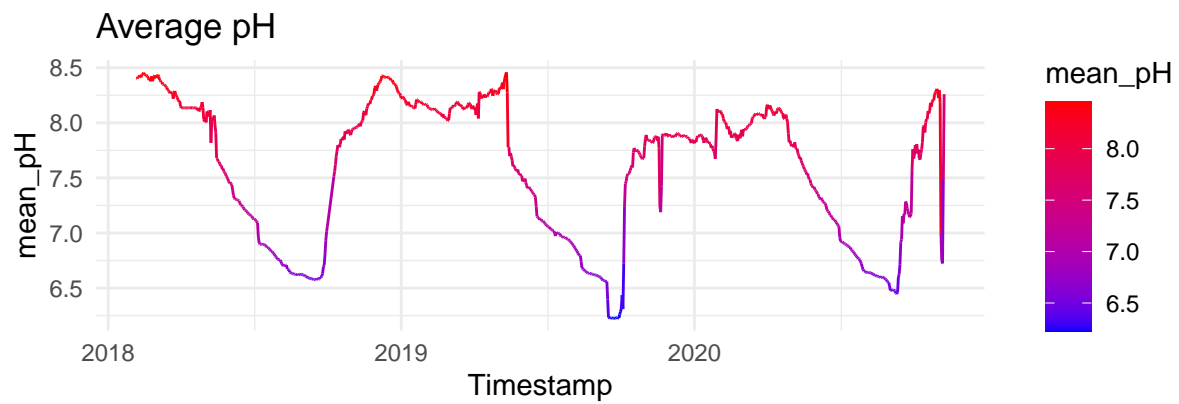
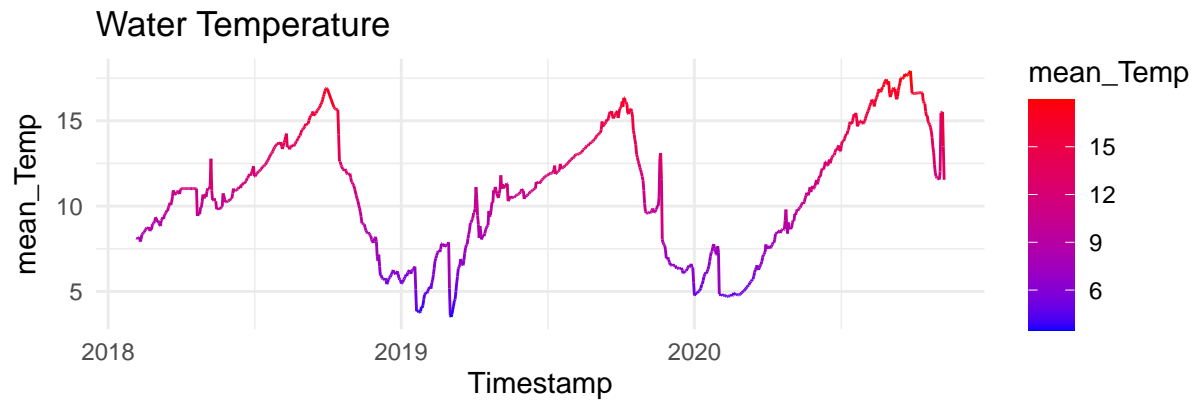
Alkalinity and hardness tend to stay fairly steady over time.



below for dosing compared to key raw water characteristics.

See





Predictor Selection

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. Conductivity is similar.

Addition of Lagged Variables

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.

Modeling

Linear Regression

```
lm_fit <- lm(Ferric.chloride..mg.L ~ ., data=lagged_data |> select(-c(Timestamp, Cationic.Polimer..mg.L,
summary(lm_fit)

##
## Call:
## lm(formula = Ferric.chloride..mg.L ~ ., data = select(lagged_data,
##      -c(Timestamp, Cationic.Polimer..mg.L)))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5702  -1.1008   0.0312   1.1992   5.1277
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -7.04725     1.58362  -4.450 9.57e-06 ***
## mean_Temp         0.08876     0.19201   0.462 0.644009
## mean_pH           1.19835     0.90382   1.326 0.185189
## mean_Conductivity -19.04679    21.77886  -0.875 0.382030
## mean_Turbidity     0.05850     0.23057   0.254 0.799775
## Temp_lag1        -0.33949     0.29773  -1.140 0.254456
## Temp_lag2         0.66989     0.19138   3.500 0.000486 ***
## pH_lag1          -1.35613     1.47784  -0.918 0.359032
## pH_lag2           3.93871     0.94008   4.190 3.04e-05 ***
## Cond_lag1         2.91313    34.37898   0.085 0.932489
## Cond_lag2         1.34539    21.76514   0.062 0.950723
## Turb_lag1        -0.07081     0.31875  -0.222 0.824251
## Turb_lag2        -0.07756     0.23055  -0.336 0.736634
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.757 on 982 degrees of freedom
## Multiple R-squared:  0.4905, Adjusted R-squared:  0.4843
## F-statistic: 78.79 on 12 and 982 DF,  p-value: < 2.2e-16
```

Interesting initial result: the pH and temperature from two days before are the only predictors with low p-values. This makes sense based on the “lag time” from water quality change to dosing change.

We will now try models incorporating only some of the predictors.

[Other Models]

Tree-Based Models

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

Classification of Dose Change

We added two variables called `dose_change_FeCl` and `dose_change_Poly` that indicate whether the chemical dose was increased, decreased, or stayed the same. Additionally, we added variables indicating change in temperature, pH, turbidity, and conductivity from the previous day.

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.