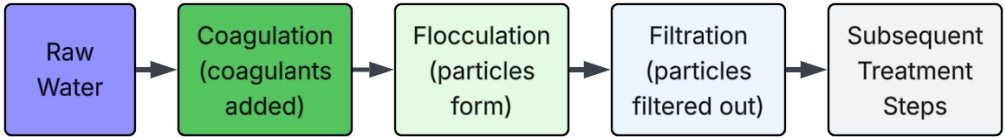
Predicting Water Treatment Plant Chemical Doses

Ethan Schilling, Joshua Tobey, Dawson Carney, Reece Carmody

11 December 2025

Background and Motivation

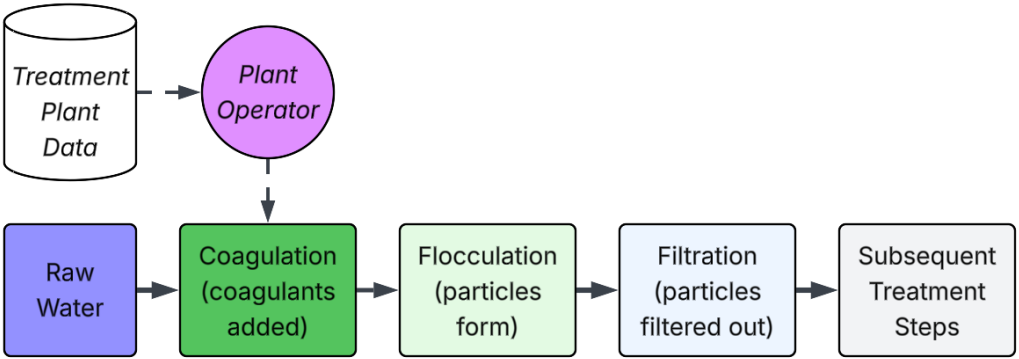
The drinking water treatment process takes water from a river, lake, reservoir, or other source, and purifies it to have it reach drinking water standards.



Background and Motivation

An operator looks at a wide array of data and makes decisions about chemical dosing.

OUR GOAL: Create a model that serves as an "operator," taking raw water data and making predictions of chemical doses.

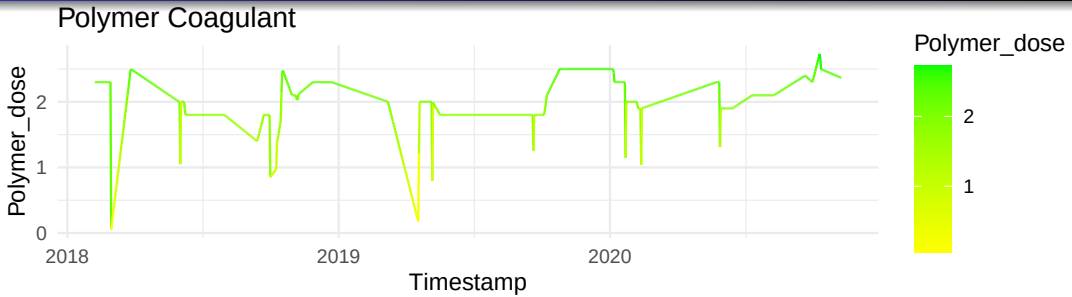


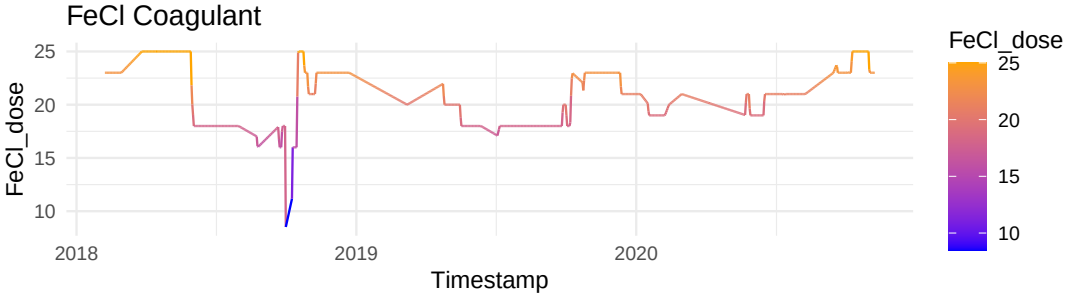
Data Overview

Three years of data from 2018-2020, provided by a Colorado water treatment plant.

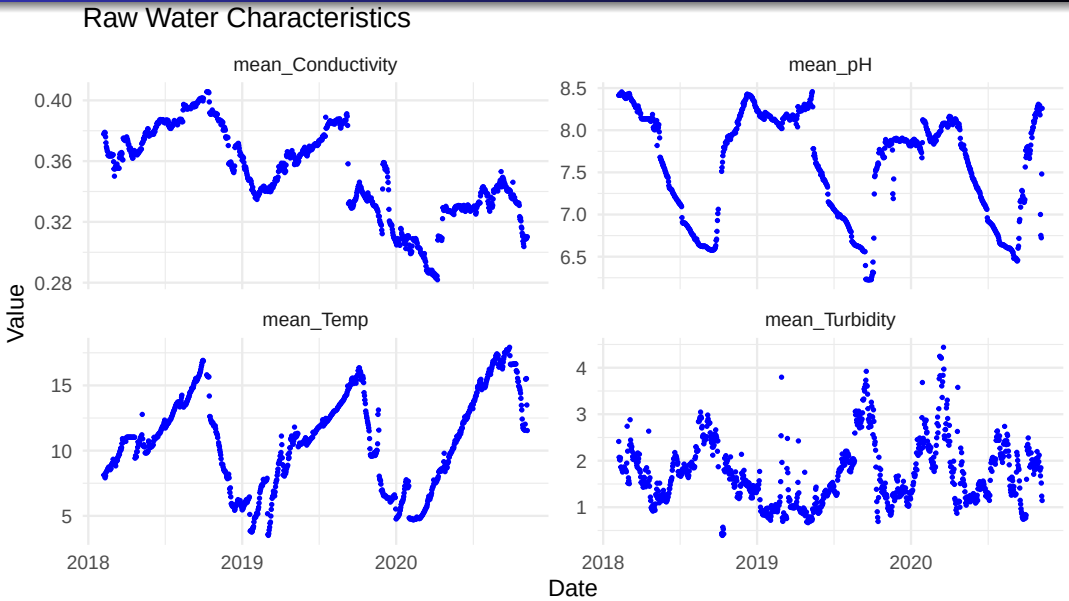
- **Raw Water Data**
  - pH
  - Temperature (of the water)
  - Turbidity
  - Conductivity
  - Suspended Grain Size (dropped)
  - Alkalinity (dropped)
  - Hardness (dropped)
- **Chemical Dosing Data**
  - Ferric Chloride Dose: primary
  - Cationic Polymer Dose: secondary
- Data Cleaning
  - Removed significant outliers
  - Took daily averages for raw water data
  - Merged common dates between datasets

Exploratory Analysis





Exploratory Analysis



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)

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

Added Variables

Introduced **lagged variables** like in the Stock dataset from the homework.

- For each of the main four variables:
  - Introduce two lagged variables
  - First represents the day before
  - Second represents two days before

Introduced a **categorical predictor** indicating whether the dose increased, decreased, or stayed the same compared to the day before.

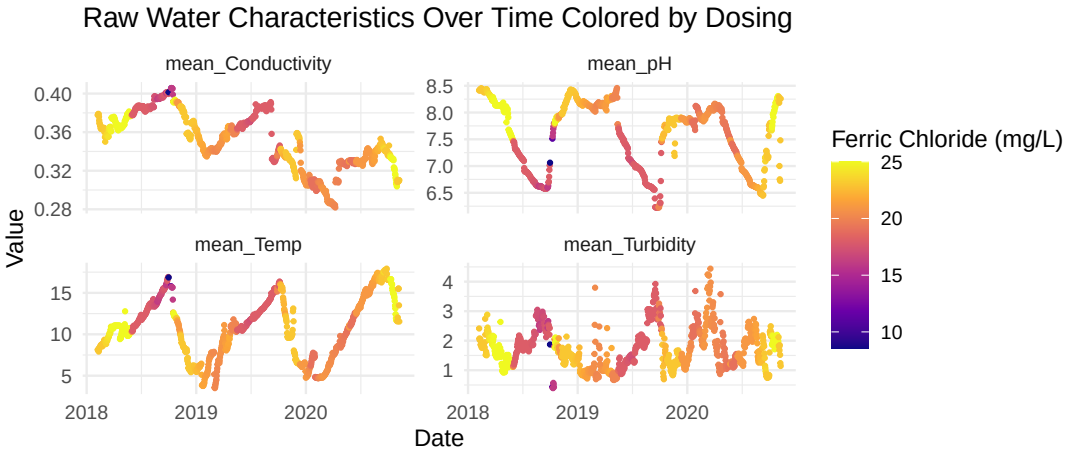
Summary of Modeling

Priorities: Interpretability, Capturing of Nonlinear Behavior

- Training and Testing
  - Training 2018-2019, Testing 2020
  - Approximately 70/30 split
- **Coagulant Dose Prediction**
  - Linear Regression Models
  - Tree-Based Models
- **Coagulant Dose Change Prediction**
  - LDA, QDA Classification
  - Multinomial Classification
  - Random Forest Classification

Coagulant Dose Prediction

GOAL: Predict coagulant doses based on raw water characteristics.



Coagulant Dose Prediction: Linear Regression

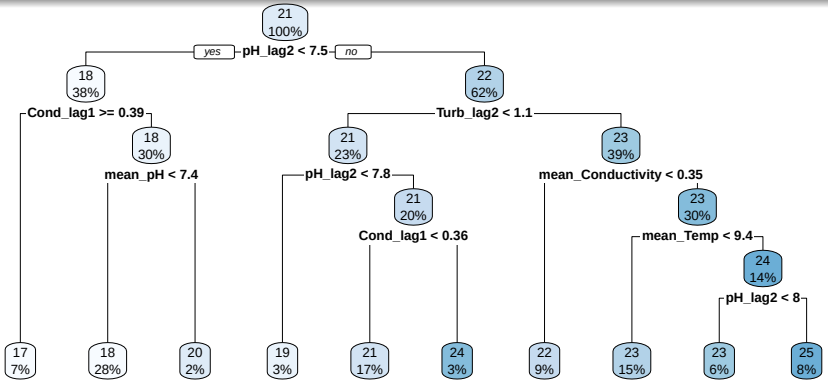
- Created 6 models with different predictors and combinations of interaction terms
- Used lag variables for temp, pH, conductivity, turbidity
- Dropped Turbidity as it was not a good predictor for FeCL dosing
- Difficult to determine what factors of the dataset are driving the predictions
- RMSE and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

Coagulant Dose Prediction: Linear Regression

Model Performance Summary

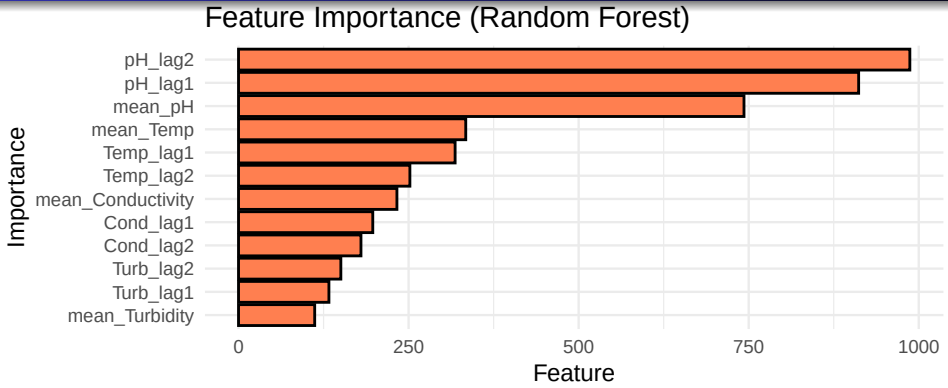
Model	RMSE	MAE	Description
Model 4	1.672	1.421	Cond × pH + Temp × Cond interactions
Model 3	1.891	1.516	Temp × Cond interaction
Model 5	2.350	1.941	Only Temp and pH
Model 0	2.971	2.580	Main effects only
Model 2	3.201	2.784	Temp × pH + Cond × Turb interactions
Model 1	3.483	2.938	Temp × pH interaction

Coagulant Dose Prediction: Decision Tree



Metric	Value
RMSE	2.756
R-Squared	0.012
MAE	2.484

Coagulant Dose Prediction: Random Forest

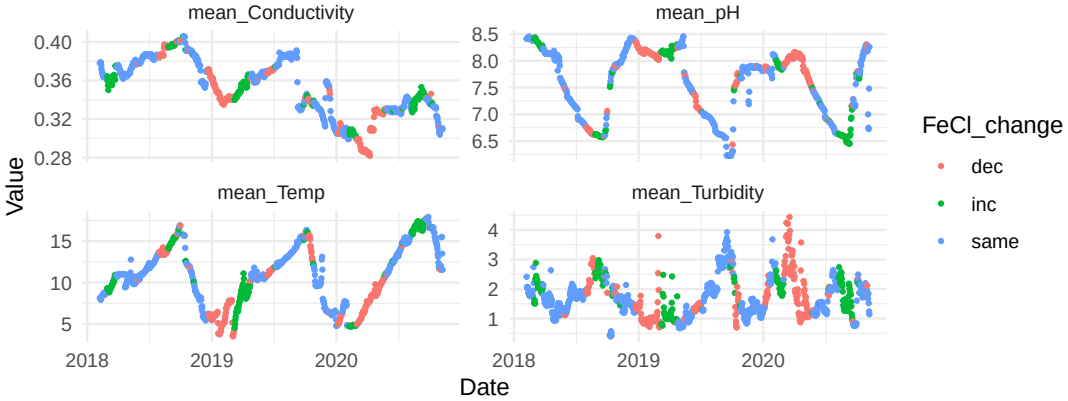


Metric	Value
RMSE	2.811
R-Squared	0.021
MAE	2.283

Coagulant Dose Change Prediction

GOAL: Predict whether coagulant dose increased, decreased, or stayed the same based on raw water characteristics.

Raw Water Characteristics Colored by Dose Change



Coagulant Dose Change Prediction: 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

Coagulant Dose Change Prediction: 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

Coagulant Dose Change Prediction: 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

Coagulant Dose Change Prediction: Random Forest

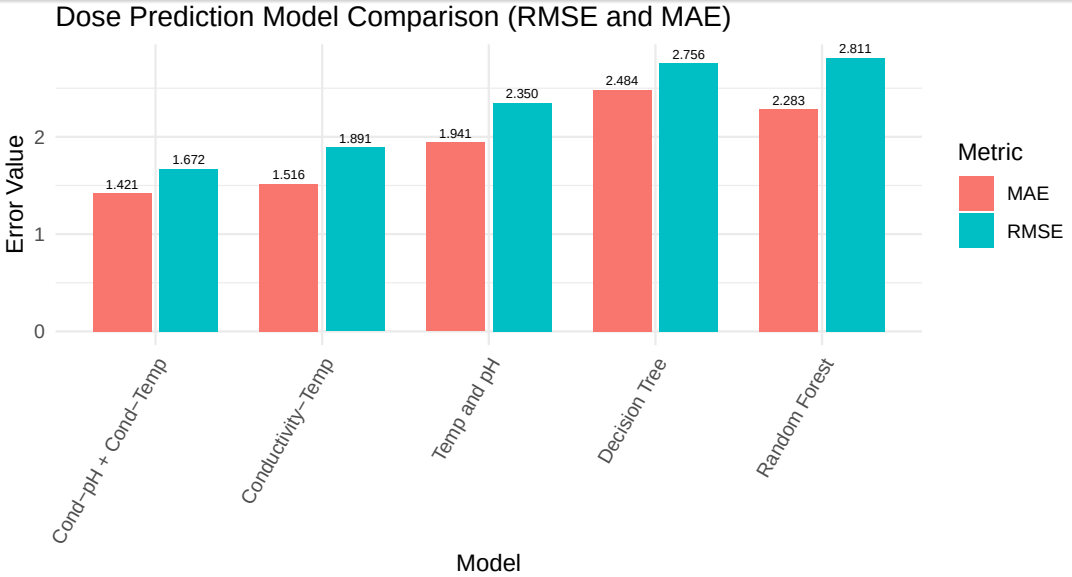
- Experimented with predictor sets and found that pH and temperature (plus the associated lagged variables) were the main effective predictors.
- Turbidity and conductivity did not improve prediction.

RANDOM FOREST ACCURACY: 39.2%

Table 6: Random Forest Confusion Matrix

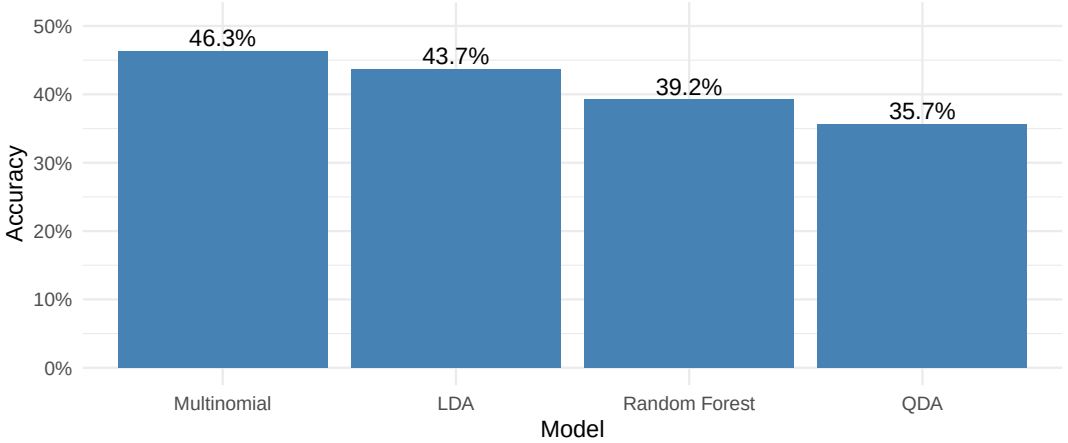
True	dec	inc	same
dec	37	6	59
inc	17	7	48
same	37	22	78

Summary of Dose Prediction Models



Summary of Dose Change Prediction Models

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



Conclusion and Next Steps

- 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