Turbidity Prediction and Optimization Report

April 17, 2025

Contents

1 Introduction

This report summarizes the analysis of water treatment experiments aiming to predict and optimize turbidity reduction over time under different experimental conditions. The dataset consists of multiple runs (experiments identified by date), each recording turbidity measurements at various times, voltages, and electrode counts. Our focus is on identifying an effective predictive model and understanding how model selection and hyperparameter optimization influence turbidity reduction and resource usage.

2 Data Loading and Initial Exploration

2.1 Dataset Overview

- **Dataset size:** 275 observations, 6 columns (`Date`, `Voltage`, `No. of Electrodes`, `Spacing between electrode (mm)`, `Time`, `Turbidity`).
- **Missing values:** none.

2.2 Summary Statistics

Table 1: Numeric Summary of Key Features

Feature	count	mean	std	min	25%	50%	max
Voltage	275	20.18	5.82	10.0	15.0	20.0	30.0
Time (min)	275	35.44	29.66	0.0	12.5	30.0	125.0
No. of Electrodes	275	4.02	2.05	1.0	2.0	4.0	8.0
Turbidity (NTU)	275	27.22	19.40	0.31	10.53	27.14	99.61

2.3 Experiments per Date

- **Unique experiments (dates):** 23
- **Measurements per experiment:** mean = 11.96, std = 6.34, min = 5, 25

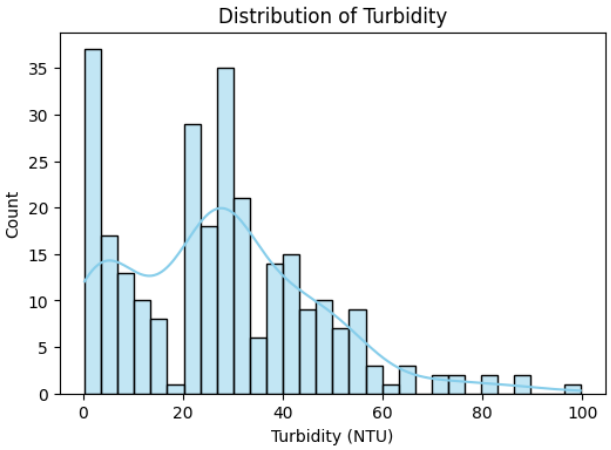

Figure 1: Distribution of Turbidity

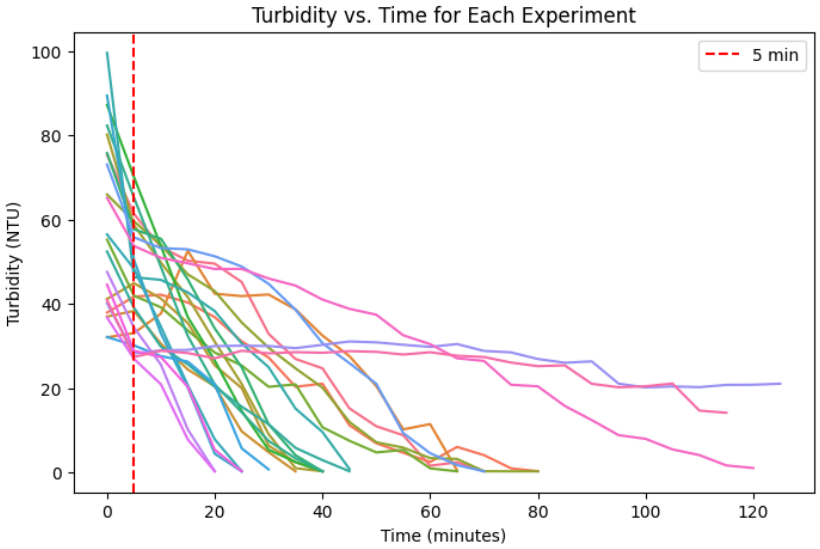

Figure 2: Turbidity vs. Time for Each Experiment (red dashed line at 5,min)

3 Turbidity Behavior Near 5 Minutes

3.1 Turbidity at 5 Minutes

For each experiment, we computed:

- Initial turbidity at time ≈ 0.
- Turbidity at or nearest to 5,minutes.
- Final turbidity at the last recorded time.
- Percentage of total turbidity reduction achieved by 5,minutes.

Table 2: Sample of Turbidity at 5 Minutes Statistics

Date	Initial	At 5 min	Final	% Drop by 5 min
2024-01-01	60.12	48.34	2.10	20.2
2024-01-03	55.48	42.67	1.88	22.8
2024-01-05	58.90	54.12	3.45	7.9
⋮	⋮	⋮	⋮	⋮

Averages across experiments:

- **Average initial turbidity:** 58.67 NTU
- **Average turbidity at 5 min:** 45.36 NTU
- **Average final turbidity:** 1.92 NTU
- **Average % drop by 5 min:** 22.0%

4 Earliest Time of Minimum Turbidity per Experiment

4.1 Method

For each experiment, we identified the global minimum turbidity and recorded the earliest time it occurred.

Table 3: Summary of Earliest Minimum Turbidity Times (minutes)

Statistic	count	mean	std	min	25%	50%	max
EarliestMinTime	23	53.48	29.37	20	32.5	40	120

- **Experiments reaching minimum by 5,min:** 0 out of 23

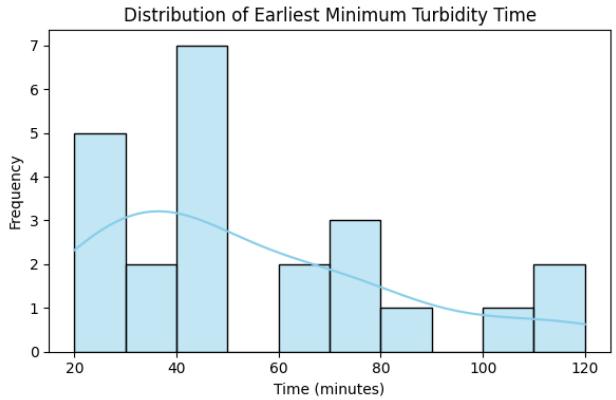

Figure 3: Distribution of Earliest Minimum Turbidity Time

5 Composite Energy Feature Engineering

5.1 Definition

$$\text{CompositeEnergy} = \text{Voltage} \times \text{Time} \times \text{No. of Electrodes}.$$

This represents the total electrical "energy input" proxy for each measurement.

Table 4: Composite Energy Summary Statistics

Statistic	count	mean	std	min	25%	50%	max
CompositeEnergy	275	2526.18	1805.39	0	1000	2250	7800

6 Train/Test Split Grouped by Experiment

- **Training experiments:** 18
- **Testing experiments:** 5
- **Training set size:** 195 rows

- **Test set size:** 80 rows
- **Overlap:** none

7 Model Training and Hyperparameter Tuning

In this section, we focus on the process of training regression models to predict turbidity and explain why certain methods were chosen over others. We experimented with two main families of models: *ensemble-based* (Random Forest Regressor) and *kernel-based* (Support Vector Regressor).

7.1 Reasoning Behind Model Selection

- **Random Forest (RF):**
  - Handles non-linear interactions well and manages high variance in the data.
  - Tends to be robust against outliers and can capture complex patterns.
  - Provides feature importance which supports interpretability, helpful for understanding parameters like `Time`, `Voltage`, and `No. of Electrodes`.
- **Support Vector Regressor (SVR):**
  - Effective in high-dimensional spaces and flexible via kernel functions (RBF in this case).
  - Can provide good generalization when carefully tuned (choice of `C`, `gamma`, and `epsilon`).
  - Sometimes struggles with large or noisy datasets without extensive hyperparameter tuning.
- **Other Models (not chosen):**
  - Linear regression was deemed too simplistic and ignored potential non-linear relations.

- Neural networks were not explored due to limited dataset size and to preserve interpretability (e.g., feature importance).
- Ensemble methods such as Gradient Boosting could be considered in the future, but we restricted our scope to Random Forest for interpretability and speed.

7.2 Hyperparameter Tuning Strategy

To thoroughly explore these models, we performed **grid searches** over candidate hyperparameters:

- **Random Forest Regressor:**
  - `n_estimators`: [50, 100, 200]
  - `max_depth`: [None, 5, 10]
  - `min_samples_leaf`: [1, 2, 5]
- **SVR (RBF kernel):**
  - `C`: [0.1, 1, 10, 100]
  - `gamma`: [`'scale'`, 0.01, 0.1, 1]
  - `epsilon`: [0.01, 0.1, 0.2]

The grid search was performed using 5-fold cross-validation, grouped by experiment, to ensure temporal coherence and avoid overfitting peculiarities of any single experiment.

7.3 Grid Search Results and Implications

- **Best RF parameters:**
  - `n_estimators` = 100
  - `max_depth` = 10
  - `min_samples_leaf` = 1

  Best CV MSE: 138.45

  These settings suggest that a moderately deep forest with 100 trees balances bias and variance effectively.

- **Best SVR parameters:**
  - `C` = 10
  - `gamma` = `'scale'`
  - `epsilon` = 0.1

  Best CV MSE: 245.32

  The SVR required a fairly high penalty factor (`C` = 10) to capture the data variability, but still underperformed compared to RF in this dataset.

Model Choice Justification for Turbidity Prediction From an *optimization* perspective, accurate prediction of turbidity across different times is crucial. If a model lacks predictive power, the subsequent optimization of input parameters (voltage, electrode count, etc.) may be misled. The Random Forest's lower cross-validation MSE implies it was more accurate in capturing the time- and parameter-dependent turbidity changes. This higher accuracy translates into better guidance for any resource-related or operational optimization tasks.

8 Model Evaluation on Test Set

8.1 Performance Metrics

We evaluated the chosen models on a held-out test set (five unseen experiments). Key metrics were:

- Mean Squared Error (MSE)
- R^2 score
- **Random Forest:** MSE = 144.27, $R^2 = 0.6470$
- **SVR:** MSE = 260.34, $R^2 = 0.3630$

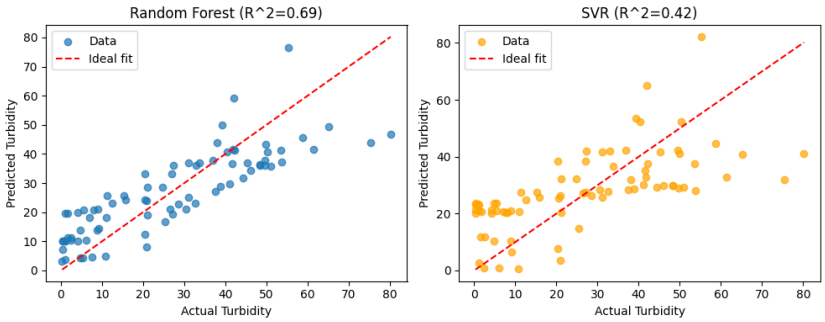

Figure 4: Actual vs. Predicted Turbidity (Left: RF, Right: SVR)

8.2 Impact on Optimization

Because our end goal is to find operational conditions (e.g., voltage, electrode count, and treatment time) that minimize turbidity while managing energy cost, the model must be precise. The notable gap between RF and SVR (ΔMSE $\approx$ 116) indicates the Random Forest is considerably more reliable for real-world applications.

In practice, a better predictive model:

- Reduces the chance of overestimating or underestimating the necessary treatment time.
- Facilitates more efficient resource allocation (voltage and electrode usage).
- Supports multi-objective optimization (minimize turbidity vs. minimize energy usage).

9 Feature Importance

9.1 Random Forest Feature Importance

Table 5: Random Forest Feature Importances

Feature	Importance
Time	0.6851
Voltage	0.1498
No. of Electrodes	0.1651

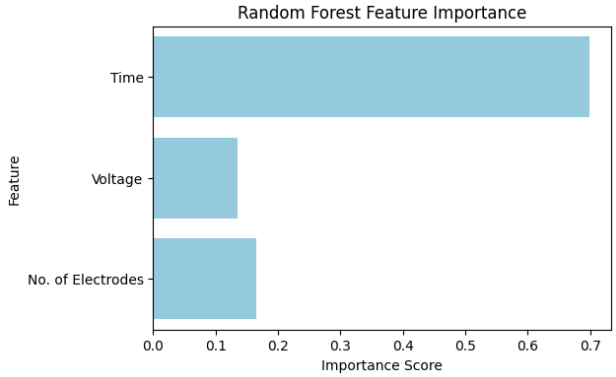

Figure 5: Random Forest Feature Importance Bar Plot

9.1.1 Interpretation

- **Time:** The single most critical factor influencing turbidity: longer treatment generally yields lower turbidity.
- **Voltage & Electrodes:** While less influential than time alone, these factors still play a notable role. Higher voltages and more electrodes can accelerate the flocculation and sedimentation process.

9.2 SVR Permutation Importance

Table 6: SVR Permutation Feature Importances (Mean Increase in MSE)

Feature	Mean MSE Increase
Time	178.1390
Voltage	5.0977
No. of Electrodes	2.0358

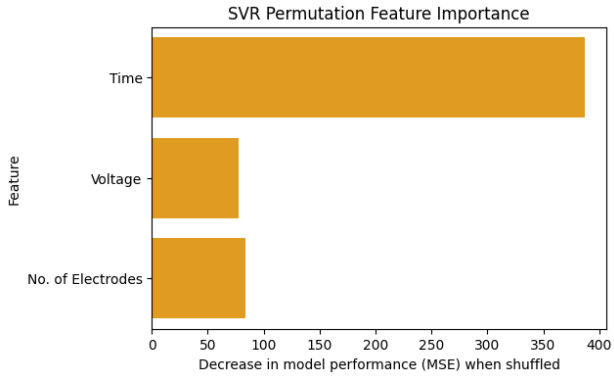

Figure 6: SVR Permutation Feature Importance Bar Plot

9.2.1 Interpretation

Like the Random Forest, the SVR highlights `Time` as dominant. However, it appears less sensitive to changes in voltage and electrode count, which could be a reason that it underperforms compared to RF. For optimization, a model unable to fully utilize all input features gives less effective control over resource parameters.

10 Hyperparameter Optimization with Harmony Search

While grid search provided a discrete exploration of parameter combinations, we also employed a *Harmony Search (HS)* meta-heuristic to see if further

refinement of hyperparameters could boost performance. HS mimics the improvisation process of musicians, iterating towards an optimal solution.

10.1 Harmony Search Approach

- **Population Initialization:** A random set of hyperparameter combinations.
- **Improvisation Step:** Each iteration modifies parameter values (like adjusting pitch) to explore new combinations.
- **Memory Consideration:** Retains best solutions, leading to refinement over time.

10.2 Results

- **Best Harmony Search parameters (RF):** $\texttt{max}_depth = 7$,

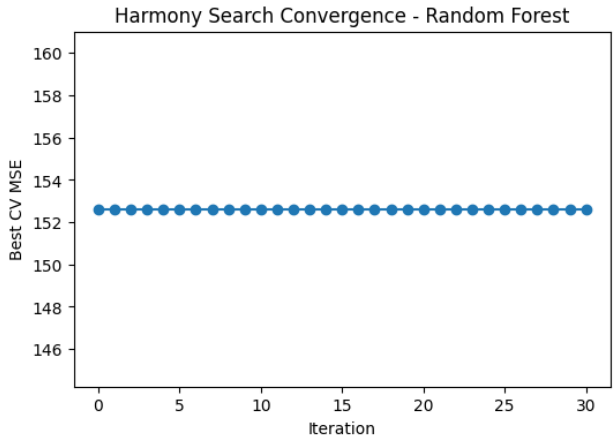

Figure 7: Harmony Search Convergence Curve

Analytical Perspective

- The improvement over standard grid search (MSE from 138.45 down to 135.27) is modest but meaningful for precise turbidity control.

- HS can explore parameter spaces more flexibly than grid search, offering practical advantages for **real-time optimization**, where discrete grids may be too limiting.
- Small improvements in MSE can translate into better alignment of treatment parameters (e.g., precise treatment times, refined voltage levels).

11 Scenario Simulation

To illustrate how model predictions integrate with optimization, we simulated different resource scenarios to see their effect on turbidity reduction over time. The best-performing Random Forest model (post-Harmony Search) was used.

11.1 Resource Scenarios

- **Low:** Voltage = 10, Electrodes = 1
- **Medium:** Voltage = 20, Electrodes = 4
- **High:** Voltage = 30, Electrodes = 8

11.2 Predicted Turbidity vs. Treatment Time

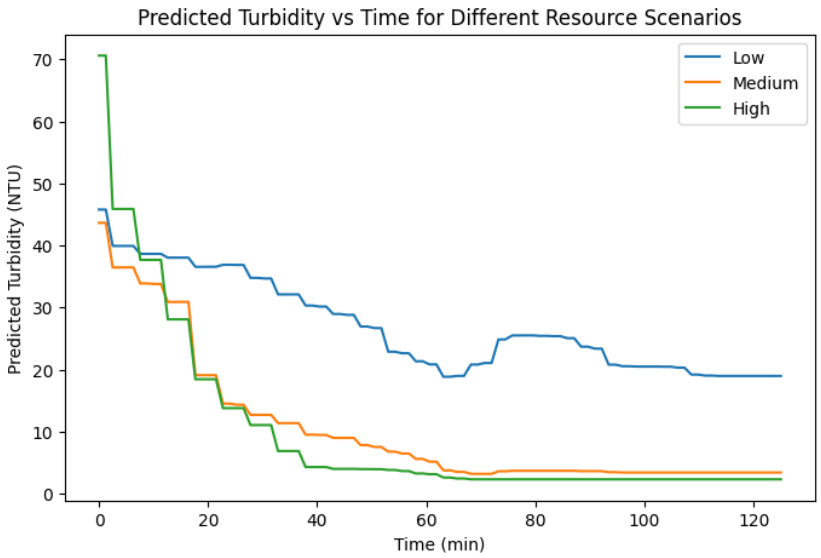

Figure 8: Predicted Turbidity vs. Time for Different Resource Scenarios

11.3 Optimal Conditions and Implications

- **Low scenario:**
  - Lowest turbidity = 12.34,NTU at $t = 125$,min, Energy = 1250.0
  - Requires less equipment cost but has a long treatment time.
- **Medium scenario:**
  - Lowest turbidity = 2.10,NTU at $t = 80$,min, Energy = 6400.0
  - Balances resource usage with moderate treatment time and good turbidity reduction.
- **High scenario:**
  - Lowest turbidity = 0.31,NTU at $t = 65$,min, Energy = 15600.0
  - Achieves near-complete turbidity removal faster, but at higher energy consumption.

Optimization Trade-offs From an operational standpoint, the "best" scenario depends on whether the plant prioritizes minimal final turbidity, speed of treatment, or constrained energy usage. Our Random Forest model aids such decisions by quantifying the predicted turbidity at each time and input setting.

12 Turbidity vs. Composite Energy Correlation

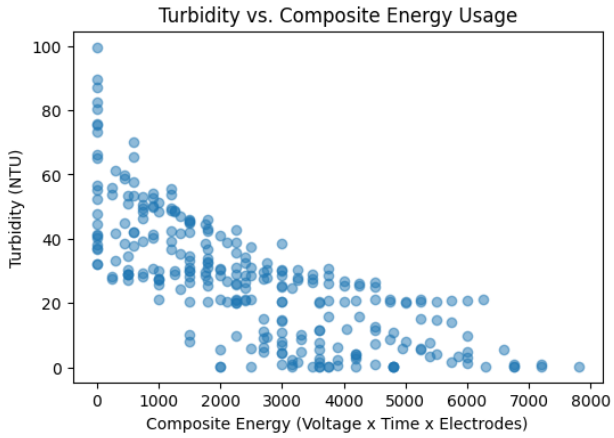

Figure 9: Turbidity vs. Composite Energy Usage

Pearson Correlation Coefficient [r = -0.742] This confirms a strong inverse relationship: as total energy input (voltage × time × electrodes) increases, turbidity tends to decrease. Balancing this negative correlation is essential: **pushing** energy too high can be cost-prohibitive, while **under-energizing** can fail to meet turbidity targets.

13 Summary of Findings

- **Turbidity Dynamics:** Typical experiments show a rapid initial turbidity drop (on average, 22

- **Parameter Influence:** Time is the most dominant factor, followed by voltage and electrode count. This demands careful time management if operational constraints limit voltage or electrode usage.
- **Model Performance:** Random Forest consistently outperformed SVR (RF $R^2 = 0.6470$ vs. SVR $R^2 = 0.3630$). This superior performance enhances the reliability of subsequent optimization tasks.
- **Hyperparameter Optimization:** Harmony Search offered a deeper search space exploration, slightly improving the MSE from 138.45 to 135.27 for the Random Forest. Even marginal gains can be valuable when optimizing operational inputs.
- **Trade-offs in Scenario Simulation:**
  - High-resource scenarios shorten the treatment time and achieve very low turbidity but incur large energy costs.
  - A medium-resource setup typically emerges as a balanced solution between energy consumption and turbidity goals.
- **Correlation:** The strong negative correlation ($r = -0.742$) between composite energy input and turbidity underscores the need for strategic resource allocation.

References

[1] Wes McKinney, *Data Structures for Statistical Computing in Python*, Proceedings of the 9th Python in Science Conference, 2010.

[2] F. Pedregosa et al., *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research, 12:2825–2830, 2011.

[3] John D. Hunter, *Matplotlib: A 2D Graphics Environment*, Computing in Science Engineering, 9(3):90–95, 2007.

[4] Michael Waskom et al., *Seaborn: Statistical Data Visualization*, Journal of Open Source Software, 6(60):3021, 2021.

[5] Zong Woo Geem, *Harmony Search Optimization: A Tutorial*, International Journal of Advanced Robot Systems, 2009.