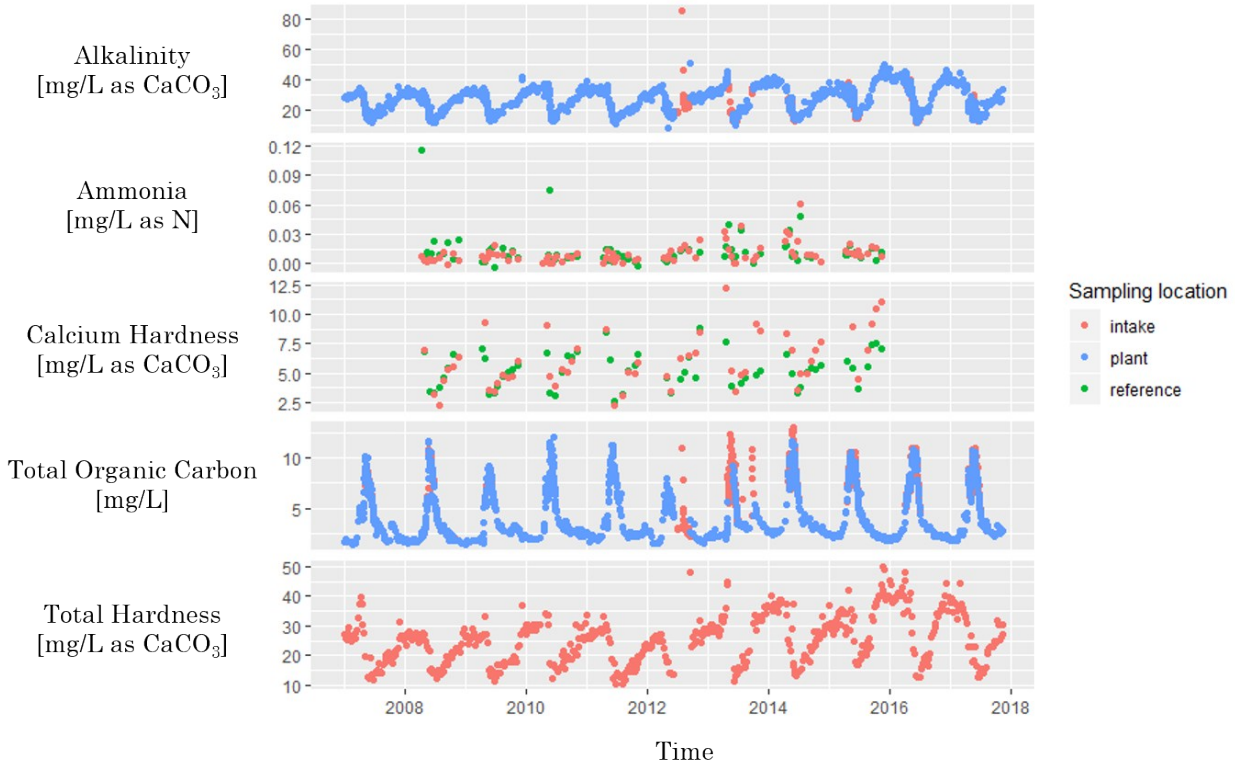
Supporting Information for "Multi-objective optimization of water treatment operations for disinfection byproduct control"

William J. Raseman[1,2]*, Joseph R. Kasprzyk[2], R. Scott Summers[2], Amanda K. Hohner[3], Fernando L. Rosario-Ortiz[2]

SI.1 Cache la Poudre water quality data

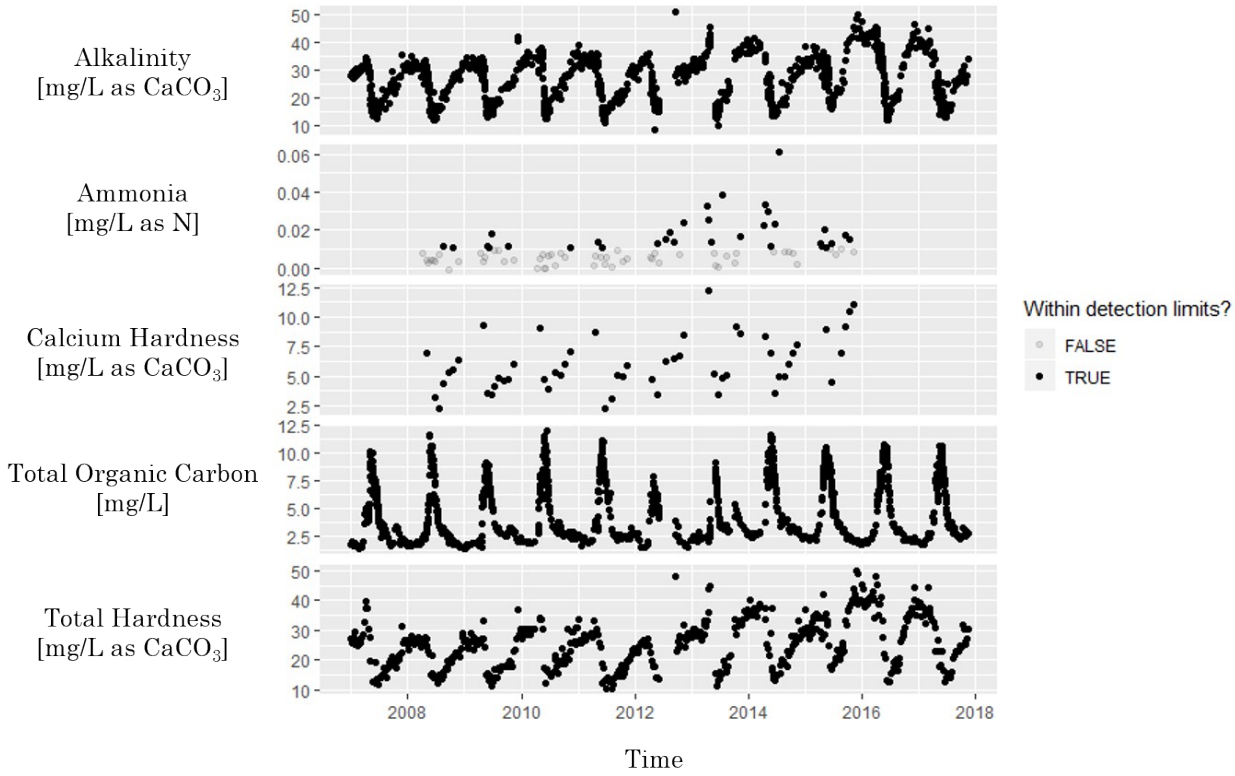
The City of Fort Collins water utility provided Cache la Poudre River water quality data shown below in **Figure SI-1**, **Figure SI-2**, and **Figure SI-3**. The "intake" sampling location refers to the intake in the Cache la Poudre River. The water is transported from the intake to the head of the treatment plant—the "plant" sampling location—via a concrete pipe, which can affect water quality. See pH and turbidity impacts in **Figure SI-3**.

In the Upper Cache la Poudre River Collaborative Water Quality Monitoring Program,[1] what we are calling the "intake" is called Poudre River above the North Fork (PNF) and "reference" is called Poudre below Rustic (PBR).



16

17 **Figure SI-1.** Alkalinity, ammonia, calcium hardness, total organic carbon, and total hardness data for the
18 intake, plant, and reference sampling points for the Cache la Poudre River (2007-2017).



19

Figure SI-2. Alkalinity, ammonia, calcium hardness, total organic carbon, and total hardness data chosen to represent influent water quality data for Cache la Poudre River (2007-2017). Whenever possible, we chose data from plant sampling location to represent the influent of the hypothetical water treatment plant for the case study; however, if plant data was of poor quality or missing, we chose to use intake data instead. After selecting representative data, we examined which data were within detection limits of measurement instruments. Since many ammonia data were below detection, we did not use these data points to generate influent water quality scenarios.

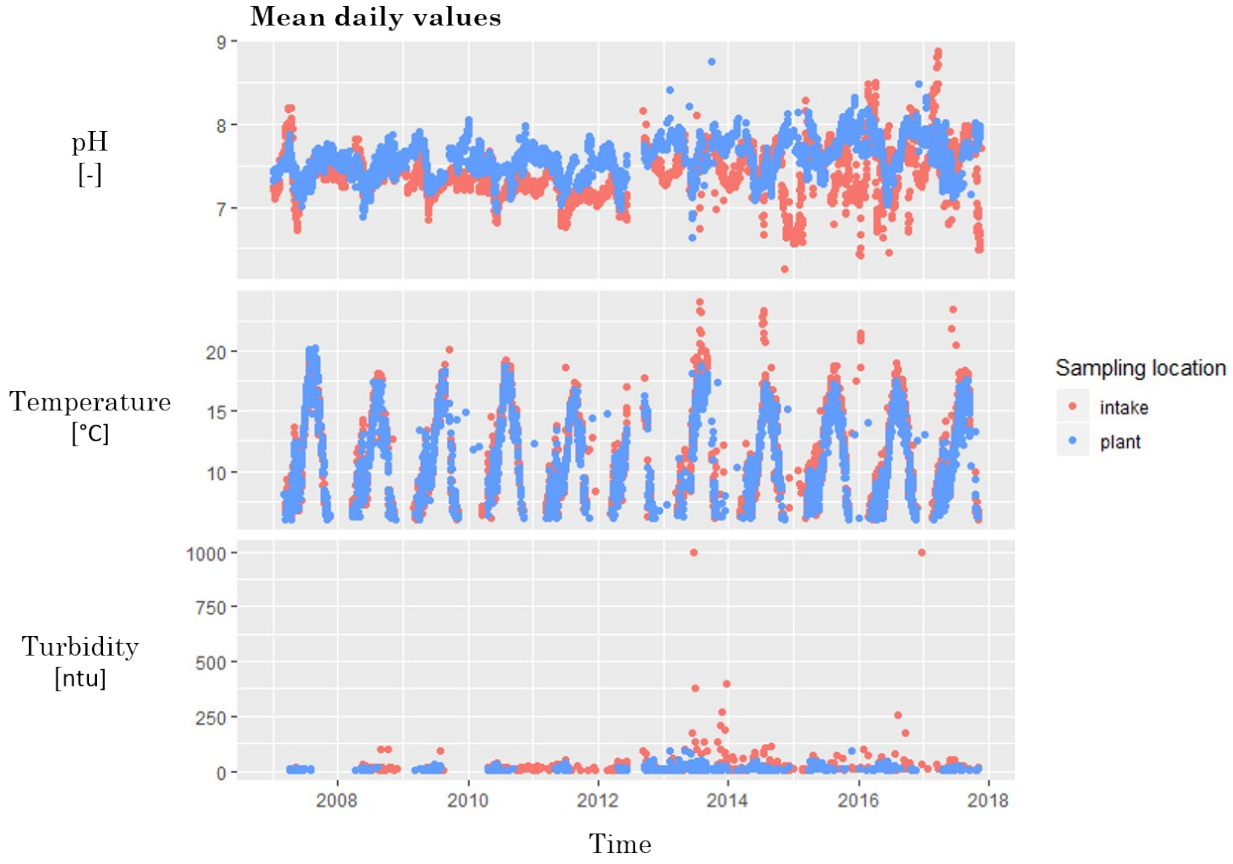


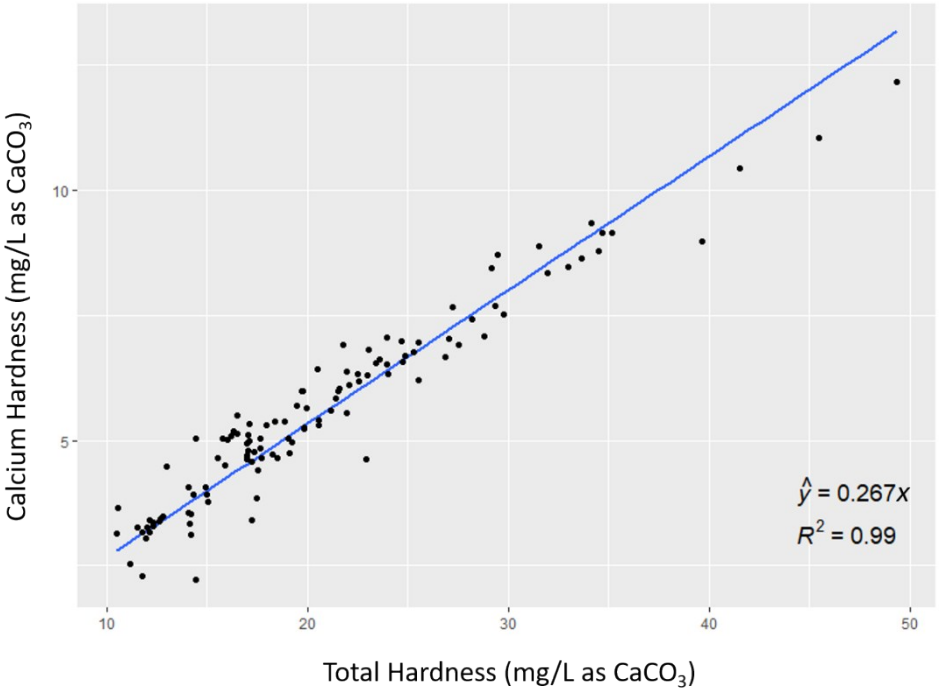
Figure SI-3. Mean daily pH, temperature, and turbidity at intake and plant sampling locations.

SI.2 Assumptions for water quality scenarios

The USEPA Water Treatment Plant Model water quality parameter inputs include pH, influent temperature, total organic carbon (TOC), UV absorbance at 254 nm (UV_{254}), bromide, alkalinity, calcium hardness, ammonia, and turbidity. Among these water quality parameters, sufficient time series data was available for TOC, alkalinity, total hardness, and turbidity for generating influent water quality scenarios as described in *SI.7 Modified k-nearest neighbor bootstrap resampling*. We estimated the remaining water quality parameters based on linear relationships between known and unknown parameters and constant concentration assumptions.

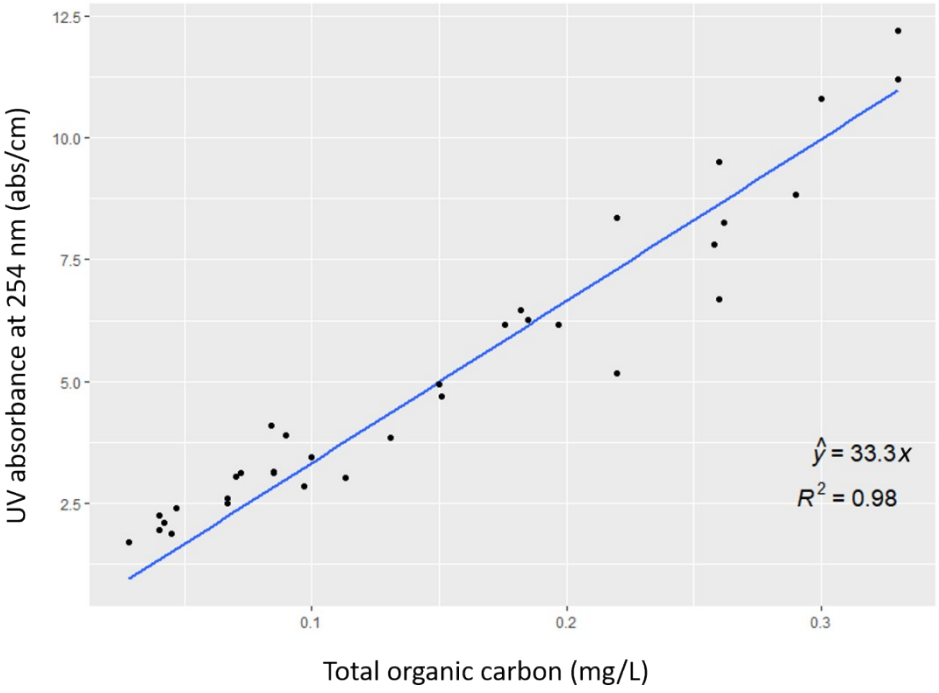
For two parameters with insufficient time series data, calcium hardness and UV_{254}, we found strong linear relationships with total hardness and TOC, respectively (**Figure SI-4** and **Figure SI-5**). These relationships allowed us to estimate calcium hardness and UV_{254} based on known parameters.

For bromide and ammonia—water quality parameters with data below detection limits—less information was available to inform an estimate of their concentrations. For bromide, in particular, there are data above detection limits. However, disinfection byproduct speciation data from treatability studies[2] found non-zero concentrations of brominated compounds, suggesting bromide concentrations were also non-zero. Because the WTP Model requires ammonia and bromide; therefore, we assumed that concentrations of these compounds were the expected value between zero and the detection limit of the measurement instruments. For instance, the bromide detect limit on record was 0.03 mg/L, therefore, a value of 0.015 mg/L was assumed. The results of sensitivity analyses and model calibration supported these assumptions (see *SI.3 Disinfection byproduct modeling and calibration* for details).



56

57 **Figure SI-4.** Linear regression (forced through the origin) of calcium hardness and total hardness which
58 includes both intake and plant Cache la Poudre River data.



59

60 **Figure SI-5.** Linear regression (forced through the origin) of ultraviolet absorbance at 254 nm and total
61 organic carbon which includes both intake and plant Cache la Poudre River data.

SI.3 Disinfection byproduct modeling and calibration

The WTP Model estimates disinfection byproduct (DBP) concentrations using multivariate power law models. The structure of these models is illustrated in the trihalomethane (THM) model for coagulated waters[3] (Equation SI-1):

$$THM = A(TOC * UV_{254})^{b}(Cl_2)^{c}(Br^-)^{d}(E)^{(pH-7.5)}(F)^{(Temp-20)}(time)^{g} \quad \text{(SI-1)}$$

Where parameters A, b, c, d, E, F, and g are empirical fitting parameters, TOC is total organic carbon (mg/L as C), UV_{254} is ultraviolet absorbance at 254 nm (1/cm), Br^- is bromide concentration (µg/L), pH is pH, $Temp$ is temperature (°C), and $time$ is the disinfection reaction time in hours.

Although the models follow this power law structure, the predictors and coefficients vary slightly between DBP models. For instance, to predict the concentration of *total* trihalomethanes (TTHM), the following coefficients are used (Equation SI-2):

$$TTHM = 23.9(TOC * UV_{254})^{0.403}(Cl_2)^{0.225}(Br^-)^{0.141}(1.1560)^{(pH-7.5)}(1.0263)^{(Temp-20)}(time)^{0.264} \quad \text{(SI-2)}$$

For review articles DBP modeling in drinking water, see Sadiq and Rodriguez[4], Chowdhury et al.[5], and Ged et al.[6].

These models were developed using bench-scale treatment tests for diverse surface waters across the United States. The WTP Model uses these models because they can capture the central tendency of DBP formation across these source waters. However, there is no guarantee that these models will provide accurate predictions on a site-specific basis. For this reason, we used bench-scale treatment data for two sampling points in the Cache la Poudre (CLP) River basin to calibrate the model for the CLP case study.

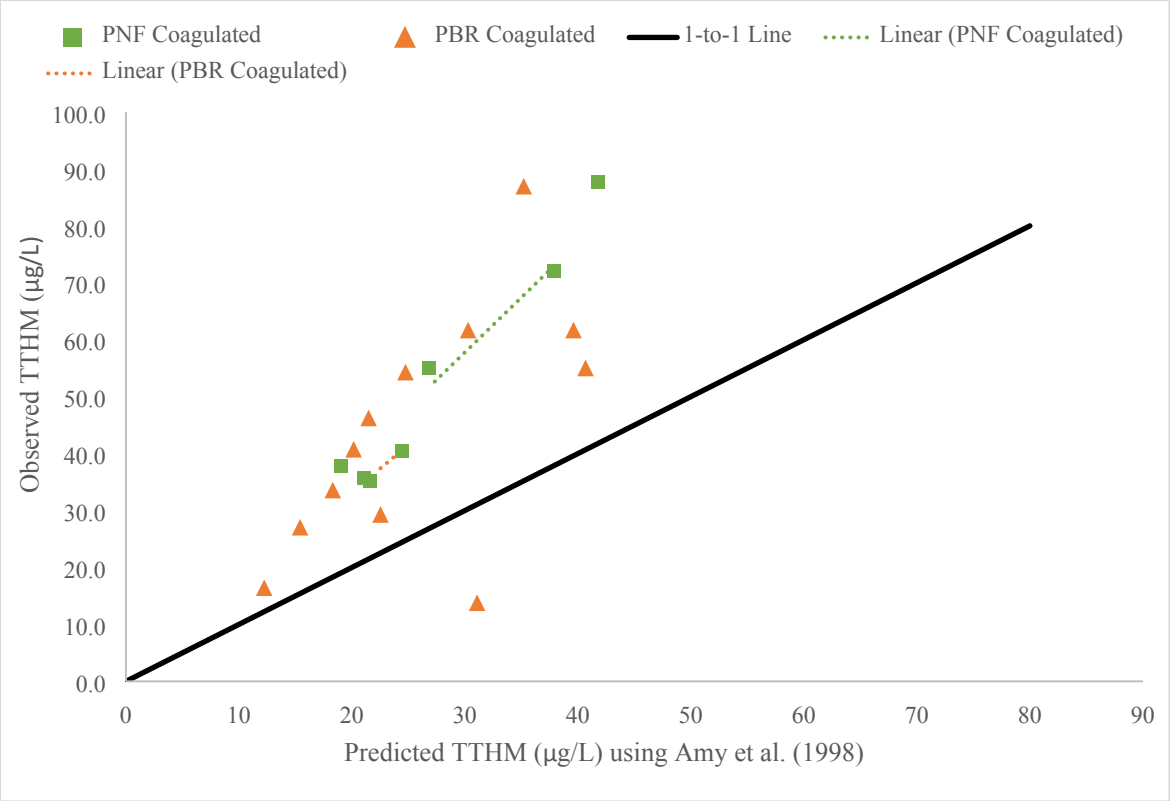
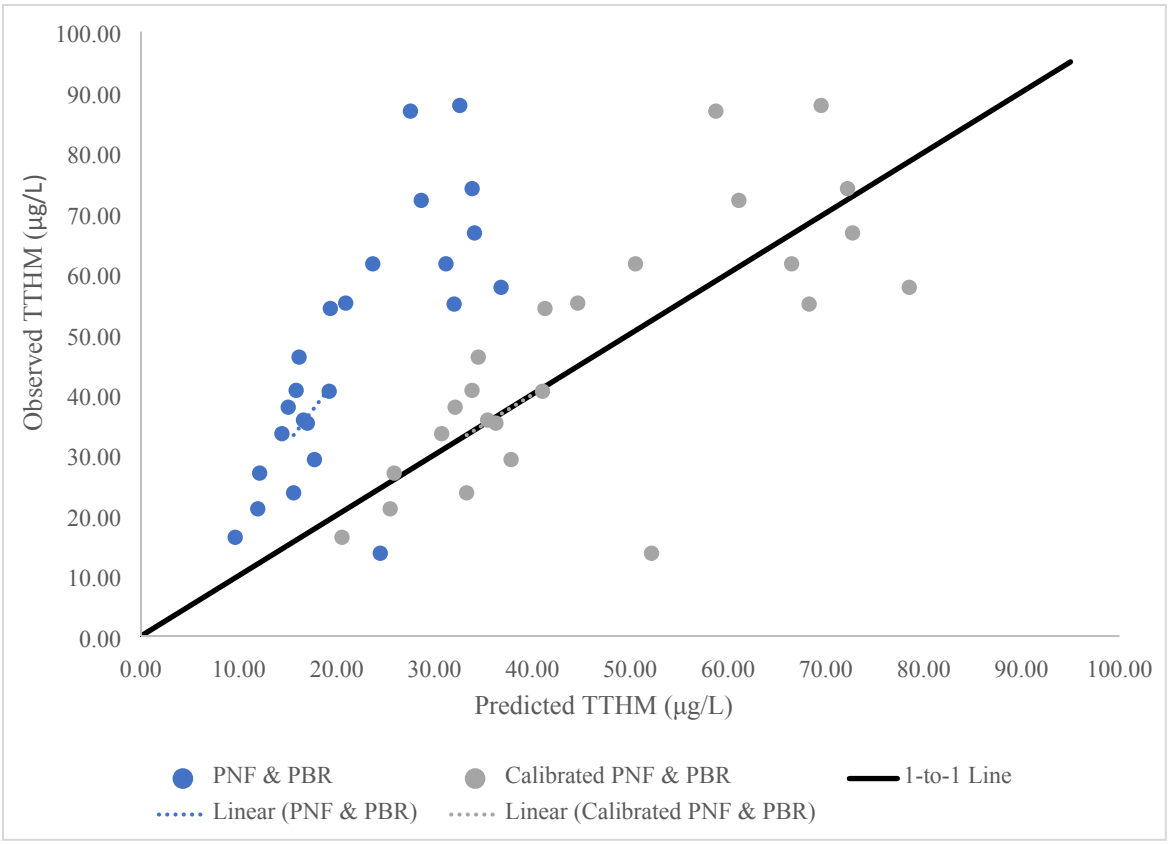


Figure SI-6. TTHM formation of observed v. estimates from the coagulated water TTHM formation model in Amy et al.[3] for burned (PNF) and unburned (PBR) sites.

To calibrate the model, we first evaluated whether DBP model predictions for the CLP River were affected by wildfire impacts. **Figure SI-6** compares the observed and predicted values at the wildfire-impacted influent (PNF) and the unimpacted reference site (PBR) and suggests similar DBP prediction between these two sites. Therefore, we decided to use both pre-wildfire and post-wildfire treatability data for the site-specific model calibration. Furthermore, since there was a limited amount of treatability data, we used both PNF and PBR data to calibrate the DBP models.

Before calibration, TTHM concentrations were underestimated by 53.4% (**Figure SI-7**), whereas the five haloacetic acids species (HAA5) were slightly overpredicted by 4.6% (**Figure SI-8**). Without changing any of the parameters in Equation SI-2, we applied a bias correction factor to the predicted DBP concentrations to fit the one-to-one line for calibration. Post-calibration R^2 for observed vs. predicted TTHM and HAA5 were 0.59 and 0.84, respectively.



95

96 **Figure SI-7.** Observed vs. uncalibrated and calibrated estimates of coagulated water TTHM formation
97 model in Amy et al.[3]

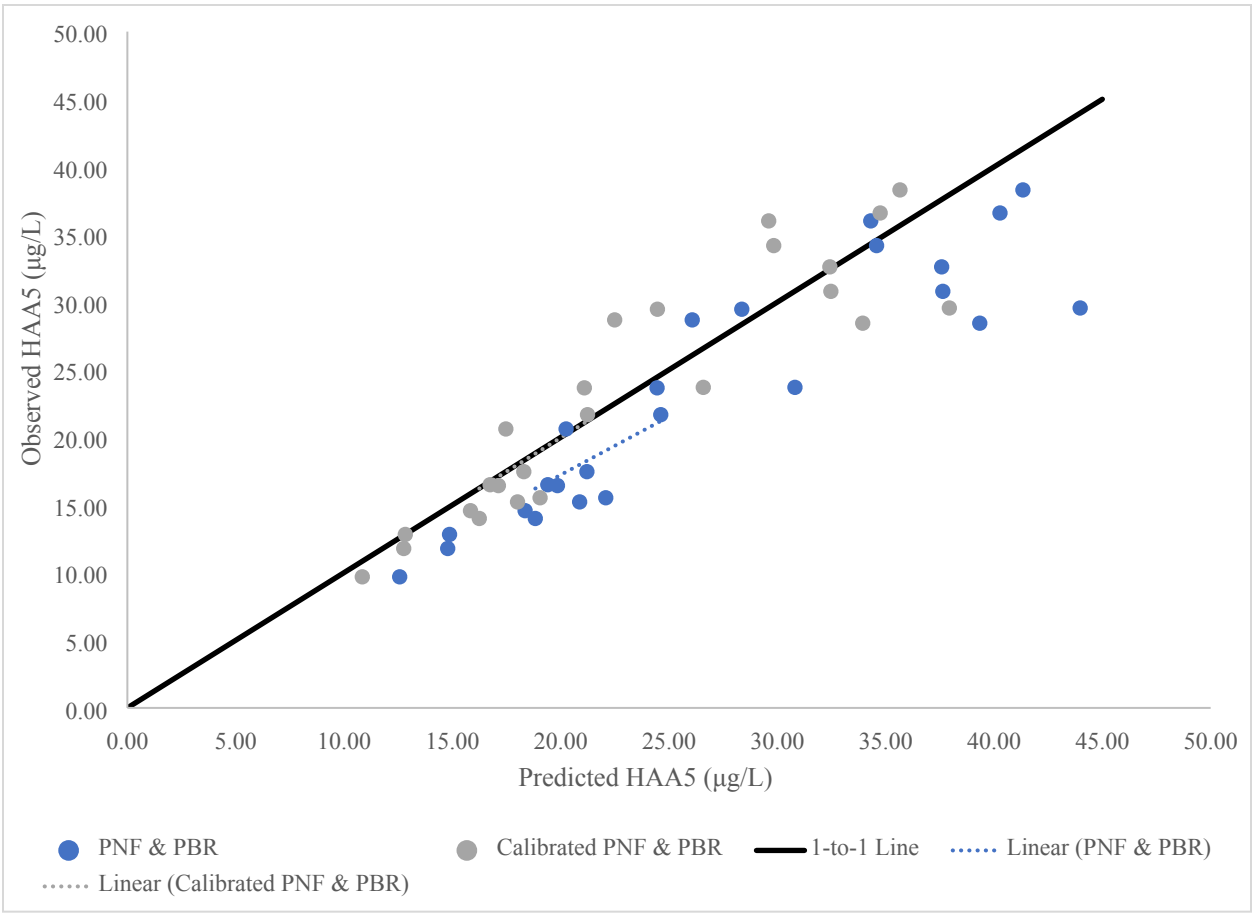


Figure SI-8. Observed vs. uncalibrated and calibrated estimates of coagulated water HAA5 formation model in Amy et al.[3]

SI.4 Modeling organic carbon removal via coagulation

Modeling the removal of organic carbon during coagulation is central to implementing the water treatment optimization problem formulation discussed in this paper. This section outlines the equations

necessary to organic carbon removal using semi-empirical sorption model developed by Edwards[7] that is used in the Water Treatment Plant (WTP) Model.

The Edwards model assumes that dissolved organic carbon (DOC) can be divided into two fractions: carbon that is complexed to a hydroxide surface (i.e., the coagulant) and carbon that cannot. These fractions are referred to as the sorbable and nonsorbable fractions, respectively (Equations SI-3 and SI-4) and the nonsorbable fraction is dependent on a linear regression of specific UV_{254} absorbance (Equation SI-5). The DOC concentration after coagulation can be calculated by substituting Equations SI-3, SI-4, SI-5, and SI-7 into a Langmuir isotherm equation (Equation SI-6). To solve for DOC, the coagulation pH (i.e., pH *after* coagulation), raw water UV_{254}, raw water DOC, and six empirical constants (K_1, K_2, x_1, x_2, x_3, and b), must be known. Coagulation pH is estimated using acid-base chemistry equations within the WTP Model and the empirical constants are listed in Table 2 of Edwards[7]. For the work in Chapter 4, the "General Al Specific" coefficients were used because aluminum sulfate (alum) was the chosen coagulant.

$$DOC_{coag} = DOC_{sorb,eq} + DOC_{nonsorb} \quad \text{(SI-3)}$$

where,

$$DOC_{nonsorb} = DOC_{raw} * F_{nonsorb} \quad \text{(SI-4)}$$

and

$$F_{nonsorb} = K_1 * SUVA_{raw} + K_2 \quad \text{(SI-5)}$$

$$\frac{DOC_{removed}}{Dose_{coag}} = \frac{a * b * DOC_{sorb,eq}}{1 + b * DOC_{sorb,eq}} \quad \text{(SI-6)}$$

where,

$$a = x_1 * pH + x_2 * pH^2 + x_3 * pH^3 \quad \text{(SI-7)}$$

(R^2_{adj} = 0.98, Standard Error of the Estimate = 0.40 mg/L, n = 608)

where,

DOC_{coag} = coagulated water DOC (mg/L): 1.0 □≤ DOC_{coag} £≤ 26

DOC_{raw} = raw water DOC (mg/L): 1.8 □≤ DOC_{raw} □≤ 26.5

$DOC_{removed}$ = [raw water DOC - coagulated water DOC] (mg/L)

$SUVA_{raw}$ = raw water SUVA (L/mg•m): 1.32 □≤ $SUVA_{raw}$ □≤ 6.11

$Dose_{coag}$ = coagulant dose (mmol Al/L): 0 □≤ $Dose_{coag}$ □≤ 1.51

pH = coagulation pH: 5.5 □≤ pH □≤ 8.0

Within the WTP Model, the Edwards model is used to predict TOC as a proxy for DOC. This TOC-based approach is supported by preliminary tests performed by Edwards that showed that using the DOC-based model for TOC had good predictive ability on independent TOC jar-test results and acceptable predictive ability for full-scale and pilot-scale test data. Edwards notes, however, that because TOC removal is nearly 100 percent during coagulation, that the DOC model typically underpredicts TOC removal. In this way, the TOC removal estimates should tend to be conservative. To refine TOC removal estimates, Tseng and Edwards[8] developed methods for calibrating the TOC model for full-scale use; however, due to data limitations this approach could not be implemented in the WTP Model, and therefore, could not be implemented in this work.

SI.5 Water quality setpoint logic and root-finding algorithm

SI.5.1 Water quality setpoint logic

Figure 4 illustrates the process flow diagram for the simulated conventional treatment plant and describes the water quality setpoints throughout the system. These values for these setpoints are determined based on either decision variables, treatment plant heuristics, or water quality regulations. Equations (SI-8 through SI-11) define the setpoint criteria for pH and alkalinity adjustment in the rapid mix unit process, corrosion control in the distribution system, and the end of distribution system chlorine residual:

$$rm_{inf,\ alk} \geq \alpha_{alk}\left(1 - \varepsilon_{tol}\right) \qquad \text{(SI-8)}$$

$$|rm_{inf,\ pH}| \geq \alpha_{pH}(1 - \varepsilon_{tol}) \quad \text{(SI-9)}$$

$$|eff_{\ pH}| \geq 8.0\,(1 - \varepsilon_{tol}) \quad \text{(SI-10)}$$

$$|eos_{cl2}| \geq 0.2\,(1 - \varepsilon_{tol}) \quad \text{(SI-11)}$$

Where $rm_{inf,\ alk}$ and $rm_{inf,\ pH}$ are the alkalinity and pH, respectively, at the influent to the rapid mix unit process. ε_{tol}, the error tolerance, is equal to 1%. $eff_{\ pH}$ is the pH at the treatment plant effluent. eos_{cl2} is the end of system chlorine residual.

The rapid mix aluminum sulfate (alum) dose is set based on the following pseudocode:

Set alum to minimum dose, 18 mg/L, to ensure turbidity removal.

Set alum to achieve Enhanced Coagulation total organic carbon removal percentage as a function of influent alkalinity[9] *plus a 5% removal safety factor to account for potential drift due to changes in pH.*

If end of system TTHM or HAA5 is greater than the maximum contaminant level (MCL), set alum based on whichever DBP MCL was not been met, attempting to achieve the DBP safety factor, β_{dbp}*, whenever possible.*

If rapid mix effluent pH is greater than or equal to pH_{min} *or the alkalinity is less than or equal to 0.0, set alum to achieve pH of* pH_{min}*. Where the minimum acceptable pH* (pH_{min}) is 5.5.

SI.5.2 Root-finding algorithm

To perform the water quality setpoint search, we developed a root-finding algorithm that uses both the bisection and secant methods [10]. We implemented this hybrid algorithm to handle the non-linearity of the multivariate power law equations used for DBP modeling and to handle the complex feedbacks in the treatment plant. The object of a root-finding algorithm is to determine the value (or values) of *x* where *f(x)*

= 0. In this case, $f(x)$ is the difference between a given water quality parameter and its setpoint target, and x is the value of the chemical dose that affects the parameter of interest.

For each iteration of the root-finding algorithm, there is an upper and lower guess, x_u and x_l, respectively. The initial guesses for each chemical dose, are listed in Table A4-1. If the signs of $f(x_u)$ and $f(x_l)$ are the same, the bisection method is used, otherwise secant method is chosen. To account for cases in which there are multiple roots, the bisection method has been modified to select the minimum, non-negative root found during the search. Non-negative roots are non-physical in this case because chemical doses cannot be negative. We choose the minimum, non-negative chemical dose because we want to minimize the cost of chemical addition. If no roots are found using the secant method, the non-negative x that corresponds to the minimum $|f(x)|$ on record is selected.

Table SI-1. Upper and lower bounds of guesses for root-finding search by chemical type

Chemical Type	Units	Lower bound	Upper bound
Lime	mg/L	0.0	1000.0
CO_2	mg/L	0.0	1000.0
Alum	mg/L	18.0	1000.0
NaOCl	mg/L	0.0	300.0

SI.6 Simulated treatment train specifications

This section describes the treatment train input parameters for the USEPA Water Treatment Plant Model. Below, t_{50} is defined as the median residence time and is typically estimated via tracer testing, and t_{10} is defined as the 10^{th} percentile residence time and is typically estimated via tracer testing.

- Influent: set based on water quality scenarios (see *SI.5 Water quality setpoint logic and root-finding algorithm*)
- Lime, Carbon Dioxide, and Alum
    - Set based on water quality setpoint logic (see *SI.5 Water quality setpoint logic and root-finding algorithm*)

- Rapid Mix
    - Volume: 0.0070 million gallons
    - Hydraulics
        - t_{50} / detention time: 1.00
        - t_{10} / detention time: 0.10
- Flocculation
    - Volume: 0.0400 million gallons
    - Hydraulics
        - t_{50} / detention time: 1.00
        - t_{10} / detention time: 0.10
- Settling Basin
    - Volume: 0.1670 million gallons
    - Hydraulics
        - t_{50} / detention time: 1.00
        - t_{10} / detention time: 0.10
- Filtration
    - Liquid volume: 0.0280 million gallons
    - Hydraulics
        - t_{50} / detention time: 1.00
        - t_{10} / detention time: 0.50
    - Chlorinated water used for filter backwash: true
    - Filter media: anthracite and sand
- Sodium Hypochlorite
    - Set based on water quality setpoint logic (see *SI.5 Water quality setpoint logic and root-finding algorithm*)
- Contact Tank
    - Volume: 1.0 million gallons
    - Hydraulics
        - t_{50} / detention time: 1.00
        - t_{10} / detention time: 0.50
- Lime and Carbon Dioxide: set based on water quality setpoint logic (see *SI.5 Water quality setpoint logic and root-finding algorithm*)
- Distribution Detention Times: average tap (1.0 day) and end of system (4.0 days)

SI.7 Modified *k*-nearest neighbor bootstrap resampling

This section describes the modified *k*-nearest neighbor bootstrap resampling algorithm developed in Raseman et al.[11]: To generate water quality scenarios, the *k*-NN method resamples historical data conditioned on a "feature vector". The choice of feature vector is dependent on two variables: the number of lags and number of water quality variables. The lag is important for realistically representing the persistence (i.e., autocorrelation, memory) of the historical water quality. For example, a source water with high organic carbon persistence means that the value tends to change very little from one timestep to the next. By considering multiple water quality variables, the *k*-NN method also attempts to capture the joint correlation among variables. Based on the feature vector for a given timestep, the algorithm calculates the nearest neighbors on the historical record from a set of candidate observations based on multivariate distance. Among the nearest neighbors, only the top *k* are considered. In this case, we use the heuristic from Lall and Sharma[12] where $k = \|\sqrt{number\ of\ neighbors}\|$. From these few neighbors, the simulated value (i.e., the successor) is selected using a weighting function that assigns the greatest probability of the 1st nearest neighbor being chosen and the least probability of choosing the *k*th neighbor. After a successor is selected, random variation is added to generate values outside of the historical record. The process then moves forward by one timestep and repeats until a user-defined number of water quality simulations is reached.

References

(1) Billica, J. A.; Loftis, J.; Moore, J. *Design of a Collaborative Water Quality Monitoring Program for the Upper Cache La Poudre River*; Prepared for the City of Fort Collins Utilities, 2008.

(2) Hohner, A. K.; Cawley, K.; Oropeza, J.; Summers, R. S.; Rosario-Ortiz, F. L. Drinking Water Treatment Response Following a Colorado Wildfire. *Water Research* **2016**, *105*, 187–198. https://doi.org/10.1016/j.watres.2016.08.034.

(3) Amy, G.; Siddiqui, M.; Ozekin, K.; Zhu, H. W.; Wang, C. Empirically Based Models for Predicting Chlorination and Ozonation By-Products: Trihalomethanes, Haloacetics Acids, Chloral Hydrate, and Bromate. August 1998.

(4) Sadiq, R.; Rodriguez, M. J. Disinfection By-Products (DBPs) in Drinking Water and Predictive Models for Their Occurrence: A Review. *Science of The Total Environment* **2004**, *321* (1), 21–46. https://doi.org/10.1016/j.scitotenv.2003.05.001.

(5) Chowdhury, S.; Champagne, P.; McLellan, P. J. Models for Predicting Disinfection Byproduct (DBP) Formation in Drinking Waters: A Chronological Review. *Science of The Total Environment* **2009**, *407* (14), 4189–4206. https://doi.org/10.1016/j.scitotenv.2009.04.006.

(6) Ged, E. C.; Chadik, P. A.; Boyer, T. H. Predictive Capability of Chlorination Disinfection Byproducts Models. *J. Environ. Manage.* **2015**, *149*, 253–262. https://doi.org/10.1016/j.jenvman.2014.10.014.

(7) Edwards, M. Predicting DOC Removal during Enhanced Coagulation. *American Water Works Association. Journal; Denver* **1997**, *89* (5), 78.

(8) Tseng, T.; Edwards, M. Predicting Full-Scale TOC Removal. *American Water Works Association. Journal; Denver* **1999**, *91* (4), 159.

(9) EPA, U. *Long Term 2 Enhanced Surface Water Treatment Rule, Toolbox Guidance Manual*; United States Environmental Protection Agency Washington, 2003.

(10) Chapra, S. C.; Canale, R. P. *Numerical Methods for Engineers*; Boston: McGraw-Hill Higher Education, 2010.

(11) Raseman, W.; Rajagopalan, B.; Kasprzyk, J.; Kleiber, W. Nearest Neighbor Time Series Bootstrap for Generating Influent Water Quality Scenarios. *Stochastic Environmental Research and Risk Assessment*.

(12) Lall, U.; Sharma, A. A Nearest Neighbor Bootstrap For Resampling Hydrologic Time Series. *Water Resources Research* **1996**, *32* (3), 679–693. https://doi.org/10.1029/95WR02966.