Predisinfection E. coli Modeling

Analysis and Prediction

2020-04-15

# Introduction

## Traditional WWTP disinfection process and control

### Regulations

* *E. coli*
  + WTD/MTD
  + Measured in the lab, results two-days later.
  + Real-time instrumentation is too costly for even the largest municipal WWTPs

### Disinfectant

* Chlorine
  + Pros: Cheap
  + Cons: Produces DBPs that are harmful to aquatic life that receive the waters and human life when drinking water sources are threatended
* PAA
  + Pros: Fewer DBP
  + Cons: Difficult to control, does not follow same simple decay kinetics as chlorine

### Control

* Flow-paced
  + Caculated dose based on mass flow rate to achieve concentration setpoint, easy to operate
  + Does not account for instantaneous chemical demand or decay changes due to changing water quality, leads to overdosing during periods of low flow
* CT-based
  + Calculates dose based on mass flow rate and retention time to achieve a CT setpoint, more difficult to program
  + Relies on fitting of first order kinetic parameters that change with time.
  + Online analyzers are costly to purchase and maintain, reactive

## Machine learning approaches to disinfection modeling

* Literature review of PAA
  + ANN modeling of chlorine dosing and residuals has been widely explored for drinking water (Sérodes, Rodriguez, and Ponton (2001)) and wastewater treatment (Gibbs et al. (2006)).
  + Wei, Haas, and Farouk (2017) was able to replicate a CFD-based PAA disinfection model (a chemically reactive time-dependent three-dimensional turbulent flow model) for a conventional 5-pass disinfection basin, using physical mixing and chemical decay kinetic parameters as ANN model inputs. The ANN was very simple, 11 inputs, 1-4 nodes in the hidden layer, and 2 outputs, fit by 40 observations and tested on 14.
* Literature review of *E. coli*
  + E.coli disinfection by chlorine and PAA follow an *n*th order kinetic reaction…
  + Most results are for total coliforms, not *E. coli*:
  + Lin et al. (2012) used an ANN to predict effluent total coliforms in a bench-scale UV disinfection system (combined with a linear model of UV dose and log inactivation) to control disinfection. However, the authors defend the strength of their model by the correlation coeffient (R2) between measured and predicted values rather than the magnitude of the error (e.g., root-mean-squared-error or RMSE). In practice, the error surrounding two of the three goal effluent total coliform counts was large (208-278 and 583-778 measured effluent total coliform counts, for the 200 and 1000 CFU/100 mL goals respectively). The third goal to achieve nondetect concentrations achieved a range of 0-25 CFU/100 mL).
  + Yu et al. (2008) used ANN to control a bench-scale chlorination-dechlorination system. 5 ANN models were built from pH and ORP in the influent, chlorination reactor and dechlorination reactor. The first model predicted the chlorine dose based on the influent quality and target effluent total coliform for feedforward control. The second model predicted effluent total coliform from the actual chlorine dose (5 minute delay from dose change predicted by first model) and influent quality to provide feedback control. A similar combination of dose-response models were built for dechlorination. Total coliform goals were acheived for a variety of reuse scenerios, but residual chlorine was much more difficult to achieve precise control.
  + Carvajal et al. (2017) investigated the realtionship between dose, time, pH, turbidity, and viral log inactivation on a bench-scale batch chlorination system for secondary treated wastewater using a Bayesian belief network (BBN). (**???**) seeded water with viruses and measured log inactivation to train Bayesian multilayer perceptron (BMLP) models to predict the required CT, one for each type of virus.
* Artificial neural networks (ANN) are a nonlinear modeling method that identifies patterns between input and output data by (1) using a large number of obervations, (2) mulitple nonlinear functions that combine different process variables, (3) trial-and-error to adjust the parameters of the nonlinear functions to accurately map input to output.
* ANN are not constructed for time-series data, i.e., correlation to previous timesteps. Recurrent neural networks (RNN) have internal memory nodes that can train based on a *sequence* of observations rather than a series of individual observations. Long short-term memory (LSTM) nodes are one example of a RNN that have been used in process data.

## Objectives

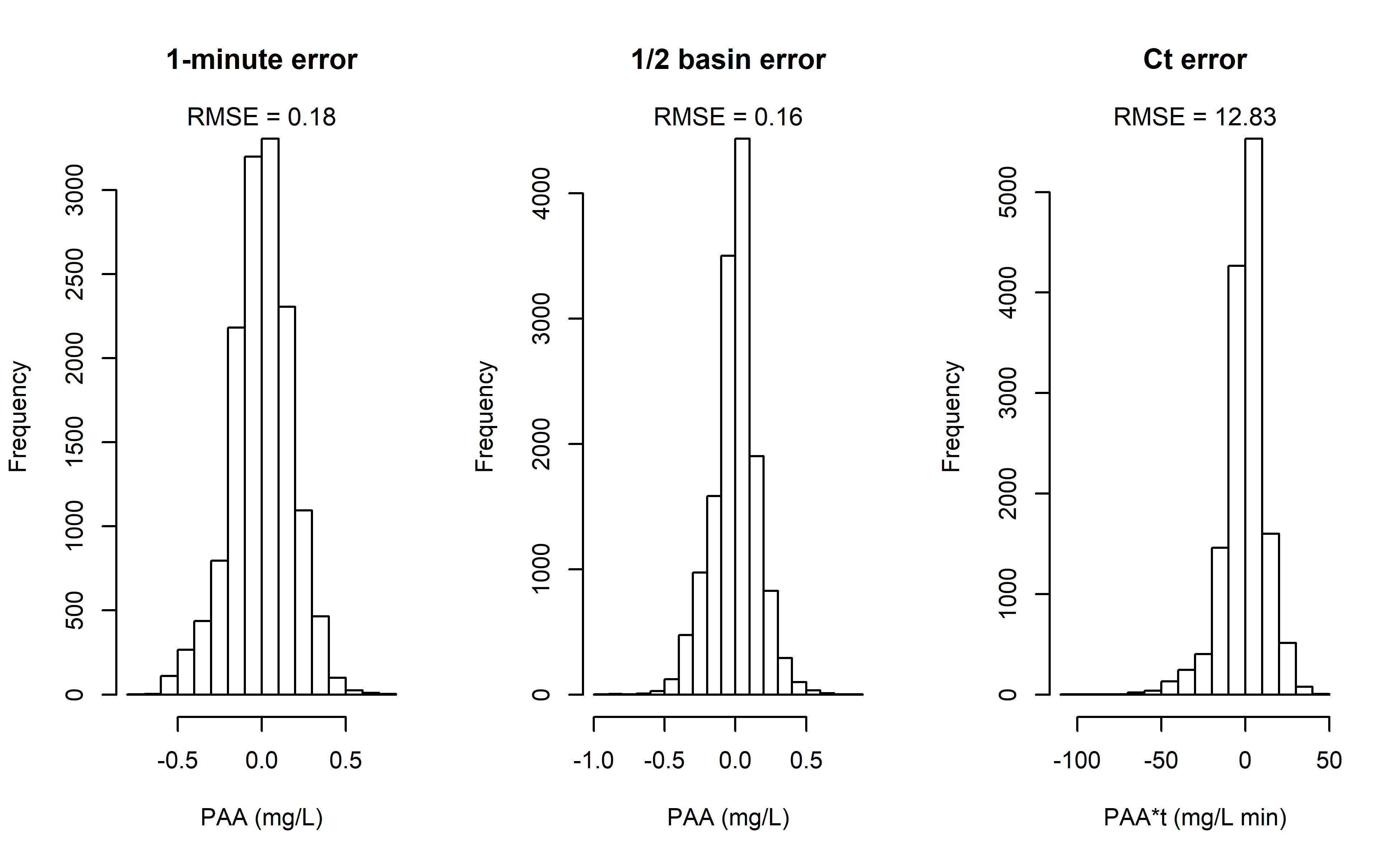
* Need a cheap, real-time measure of PAA and E.coli concentrations for accurate disinfection control in WWTP
  + Need to understand the environmental and operational conditions that impact (1) PAA demand and decay and (2) pre- and post-disinfection E.coli.
  + Need to use unconventional modeling approaches, such as RNN, to (1) estimate first order decay parameters for PAA and (2) model *E. coli* removal

# Materials and Methods

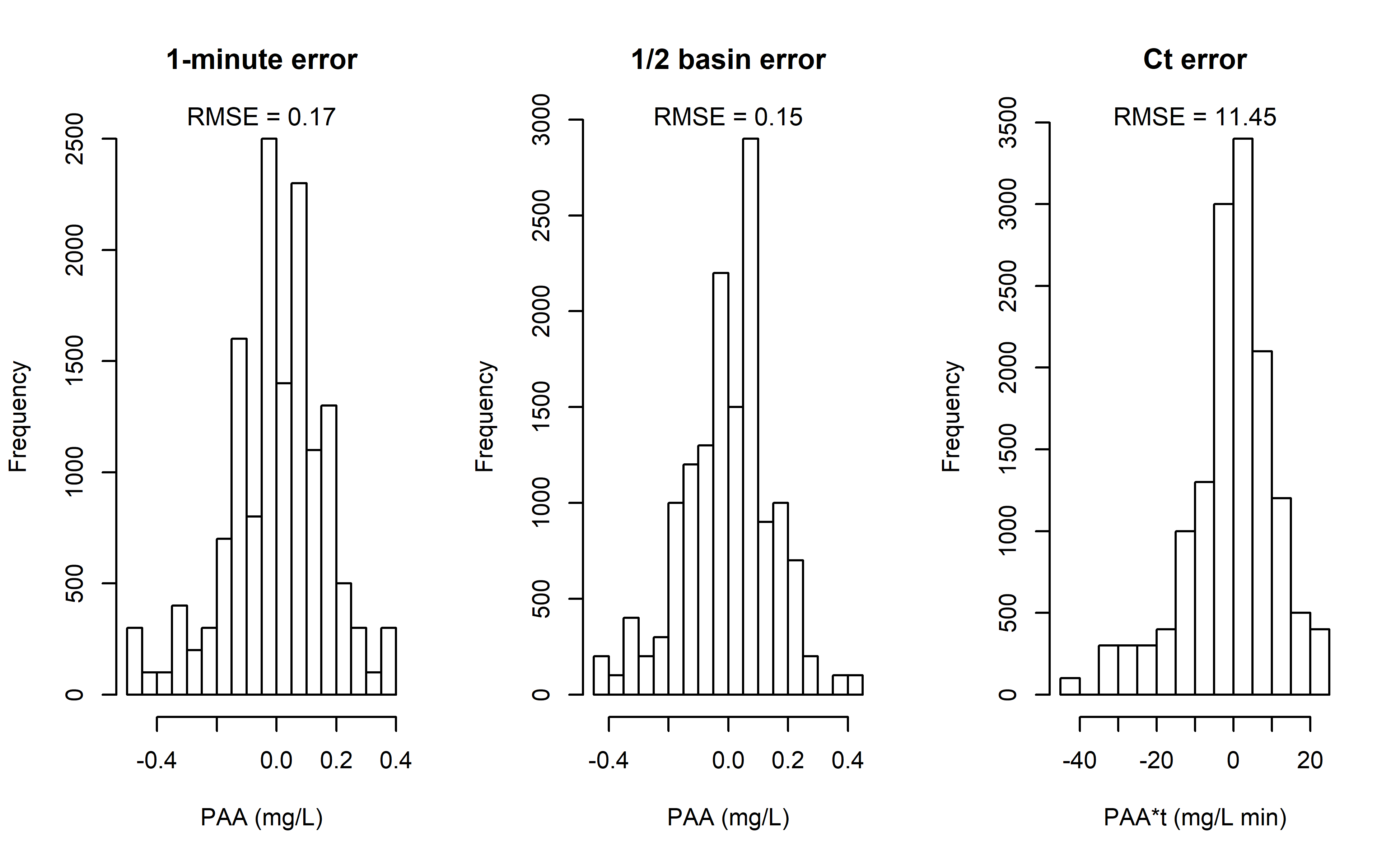
## RWHTP

* Two separate treatment trains: North and South
* Data collected:
  + PAA profiles from X-Y
  + Daily E. coli measurements from X-Y
  + Online instrumentation of upstream treatment processes
  + Daily-weekly samples of upstream treatment processes, both instantaneous grab and 24-hour flow-weighted composites

| **North Variable** | **Collection Method** |
| --- | --- |
| Pump Flow Based PAA Dose | Online |
| PAA 1 min Sample | Grab |
| PAA 1 2 Basin Sampling | Grab |
| N Eff TSS Conc | Online |
| Temp of NSEC Main Ch | Online |
| Temp of the Atmos | Online |
| Secondary Effluent Flow | Online |
| N Basin Outfall | Online |
| CODto mg L | Online |
| TSS mg L | Online |
| UVT | Online |
| CODds mg L | Online |
| SACto 1 m | Online |
| NSEC Aerobic SRT | Online |
| NSEC Effluent NH3 | Online |
| NSEC Effluent NO3 | Online |
| NSEC Effluent OP | Online |
| NSEC Effluent TSS | Online |
| NSEC Effluent NO5 | Online |
| NSEC Effluent Flow | Online |



## Without visual data



### *E. coli* model input data

Data from the North and South that are used to predict *E. coli* are summaried below in Tables 2 and 3. Figures 1 and 2 show the predisinfection *E. coli* and the log10-transformed predisinfection *E. coli* as a timeseries and boxplot. The boxplot illustrates that the raw predisinfection *E. coli* is heavily skewed (i.e., many outliers on one side of the median). By using a log10-transformation, the data is more normally distributed.

**Table 2.** Variables monitored in the North to predict E. coli performance. Number of observations is calculated by the number of variables that have a value collected within 24-hours prior to the E.coli sample

**Figure 1.** North predisinfection *E. coli* for training (before purple line) and testing (after purple line) period. A log10-transform approximates normally distributed training and testing data.

**Table 3.** Variables monitored in the South to predict predisinfection *E. coli* performance. Number of observations is calculated by the number of variables that have a value collected within 24-hours prior to the *E. coli* sample

**Figure 2.** South predisinfection *E. coli* for training (before purple line) and testing (after purple line) period. A log10-transform approximates normally distributed training and testing data.

## Recurrent neural networks

* First 90% of data is used to train a recurrent nerual network (RNN)
* Last 10% of data is used to test RNN
* Single layer of *n* long short-term memory (LSTM) nodes, where *n* is the number of variables in the model

# Results

## *E. coli* prediction

### Predisinfection *E. coli*

A few data transformations were used to improve model fit: \* average process data to better represent *E. coli* sampling conditions, ranging from 15-minutes to 24-hours \* log10-transform of *E. coli* \* lag *E. coli* by 2-days While *E. coli* is more highly correlated to the previous *E. coli* observation (1-day), the average time from when a sample is taken to a result is 2-days. Thus, to simulate the use of this model in “real-time” a lag of 2 observations was used.

A common problem in machine learning is *overfitting*. Overfitting occurs when the model-building method continues to adjust model parameters beyond the general trends and instead fits to error within the training dataset. This causes a deceptively high training R2, as the testing RMSE is also very high. RMSE can be interpreted as the standard deviation of the unexplained variance and is used to measure testing error here. A good model for real-time prediction minimizes RMSE. Thus, here we examined different number of training periods, from 100-2000. Figure 3 illustrates how R2 is not a good predictor of test RMSE. Currently, *E. coli* is ‘predicted’ by assuming the last known value approximates current conditions. In modeling, this is refered to as *persistence* and is represented as the red, dashed line in Figure 3. Because the model error is less than persistence for 100-750 epochs, and minimized for 100-500 epochs, the ideal number of training epochs being between 100-500.

**Figure 3.** South predisinfection *E. coli* is predicted using a simple RNN, using a various number of epoch (e.g., “re-train” parameter estimations). The red line is the *persistence* RMSE, which is the RMSE calculated from the previously known *E. coli* value. Compared to the model predicted RMSE (black circles), 100-500 training epochs improve upon the current method of estimating *E. coli* (being less than the persistance prediction).

In Figure 4, the best performing model predictions (lowest test RMSE) used are plotted alongside the actual *E. coli* for the entire testing period. For the majority of the testing period, the errors are relatively low, exept for an event on February 25, 2020.

### Log removal

### Sensitivity analysis

Carvajal, Guido, David J. Roser, Scott A. Sisson, Alexandra Keegan, and Stuart J. Khan. 2017. “Bayesian Belief Network Modelling of Chlorine Disinfection for Human Pathogenic Viruses in Municipal Wastewater.” *Water Research* 109 (February): 144–54. <https://doi.org/10.1016/j.watres.2016.11.008>.

Gibbs, M. S., N. Morgan, H. R. Maier, G. C. Dandy, J. B. Nixon, and M. Holmes. 2006. “Investigation into the Relationship Between Chlorine Decay and Water Distribution Parameters Using Data Driven Methods.” *Mathematical and Computer Modelling*, Application of Natural Computing Methods to Water Resources and Environmental Modelling, 44 (5): 485–98. <https://doi.org/10.1016/j.mcm.2006.01.007>.

Lin, Chuang-Hung, Ruey-Fang Yu, Wen-Po Cheng, and Chun-Ru Liu. 2012. “Monitoring and Control of UV and UV-TiO2 Disinfections for Municipal Wastewater Reclamation Using Artificial Neural Networks.” *Journal of Hazardous Materials* 209-210 (March): 348–54. <https://doi.org/10.1016/j.jhazmat.2012.01.029>.

Sérodes, Jean-B, Manuel J Rodriguez, and Annie Ponton. 2001. “Chlorcast: A Methodology for Developing Decision-Making Tools for Chlorine Disinfection Control.” *Environmental Modelling & Software* 16 (1): 53–62. <https://doi.org/10.1016/S1364-8152(00)00064-5>.

Wei, Wangshu, Charles N. Haas, and Bakhtier Farouk. 2017. “Optimized Design of Wastewater Disinfection Reactors Based on an Artificial Neural Network Metamodel.” In *ASME 2016 International Mechanical Engineering Congress and Exposition*. American Society of Mechanical Engineers Digital Collection. <https://doi.org/10.1115/IMECE2016-65139>.

Yu, Ruey-Fang, Ho-Wen Chen, Wen-Po Cheng, and Yu-Chiu Shen. 2008. “Dynamic Control of Disinfection for Wastewater Reuse Applying ORP/pH Monitoring and Artificial Neural Networks.” *Resources, Conservation and Recycling* 52 (8): 1015–21. <https://doi.org/10.1016/j.resconrec.2008.03.007>.