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 *E. coli*
  + E.coli disinfection by chlorine and PAA follow an *n*th order kinetic reaction…
  + Some literature exists, but with 5-10 years of data and never implemented in real-time
  + 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.
  + Gibbs et al. (2006) used ANN to model disinfection in water treatment.
* 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

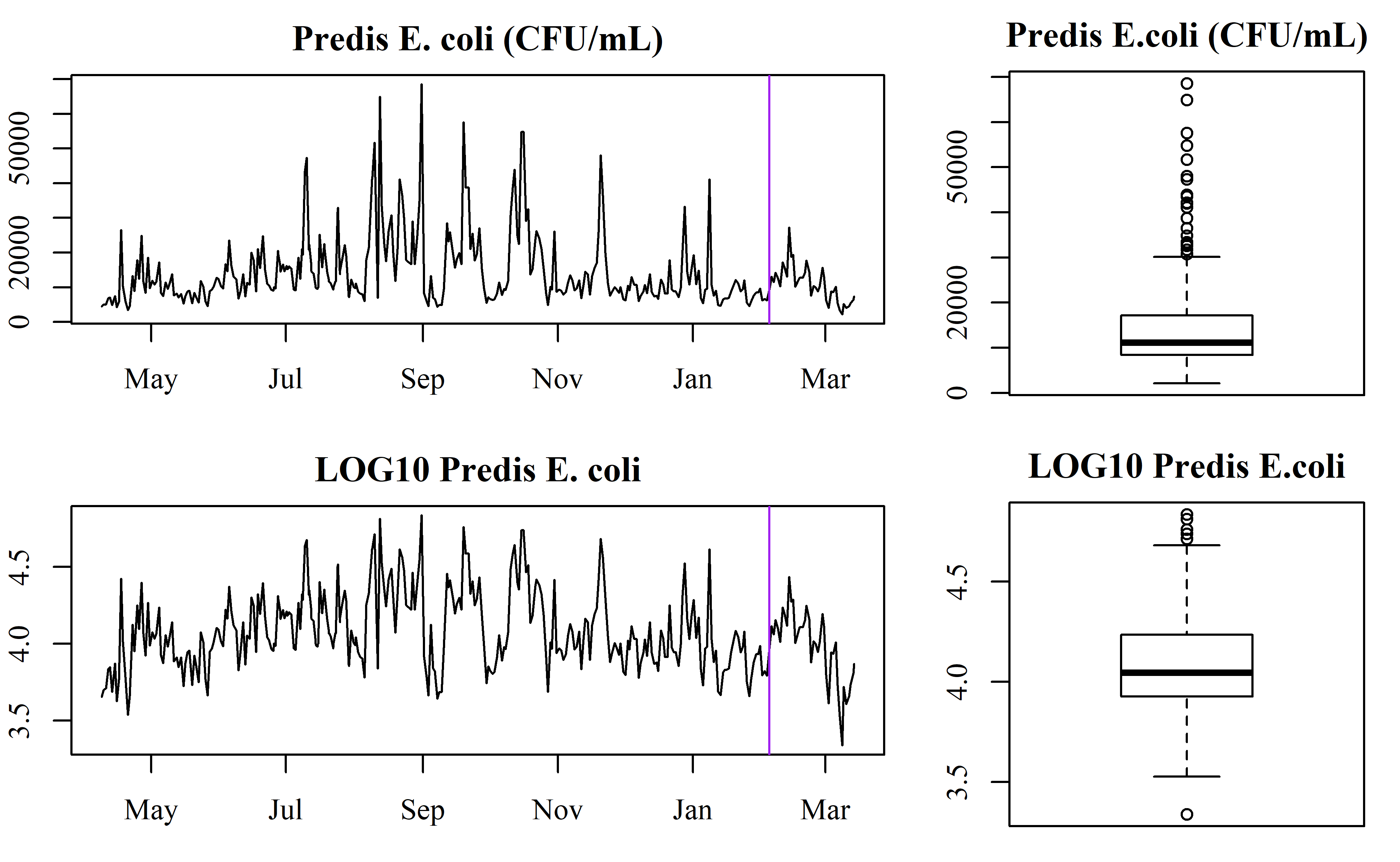
### *E. coli* model input data

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

| **North Variable** | **Collection Method** | **Number of Observations** |
| --- | --- | --- |
| North Nitrification Effluent TSS | FC24 | 174 |
| NSEC Quad 1 MLR SVI | Grab | 153 |
| NSEC Quad 2 MLR SVI | Grab | 160 |
| NSEC Quad 3 MLR SVI | Grab | 154 |
| NSEC Quad 4 MLR SVI | Grab | 133 |
| AC N94A | Online | 385 |
| AC N94B | Online | 385 |
| AC N94C | Online | 385 |
| AI N92A | Online | 385 |
| AI N92C | Online | 385 |
| AI N92D | Online | 385 |
| AI N92H | Online | 385 |
| AI N93D | Online | 385 |
| AI N99C | Online | 385 |
| ALk | FC24 | 103 |
| ASRT ASRT N | Online | 385 |
| BOD | FC24 | 171 |
| CBOD | FC24 | 161 |
| COD | FC24 | 60 |
| ECIDX | Grab | 385 |
| FC N231 | Online | 385 |
| FC N236 | Online | 385 |
| FI T631 GTE to SAR 2 Flow | Online | 385 |
| FI T632 GTE to SAR 4 Flow | Online | 385 |
| Main Inf Channel NSEC TI N171 | Online | 385 |
| NH3A | FC24 | 337 |
| NO5 | FC24 | 132 |
| NSEC EFF FLOW FY F225 | Online | 385 |
| NSEC INF FY F25 | Online | 385 |
| OP | FC24 | 41 |
| PW | FC24 | 128 |
| Quad 1 Ave Blanket Depth NSEC LI N561Q | Online | 384 |
| Quad 1 Basins In Service | Online | 385 |
| Quad 2 Basins In Service | Online | 385 |
| Quad 3 Basins In Service | Online | 385 |
| Quad 4 Basins In Service | Online | 385 |
| RAS %AE Basin Inf NSEC FC N200B2 | Online | 384 |
| TI R3003 | Online | 383 |
| TIN | FC24 | 123 |
| TKNH | FC24 | 132 |
| TN | FC24 | 122 |

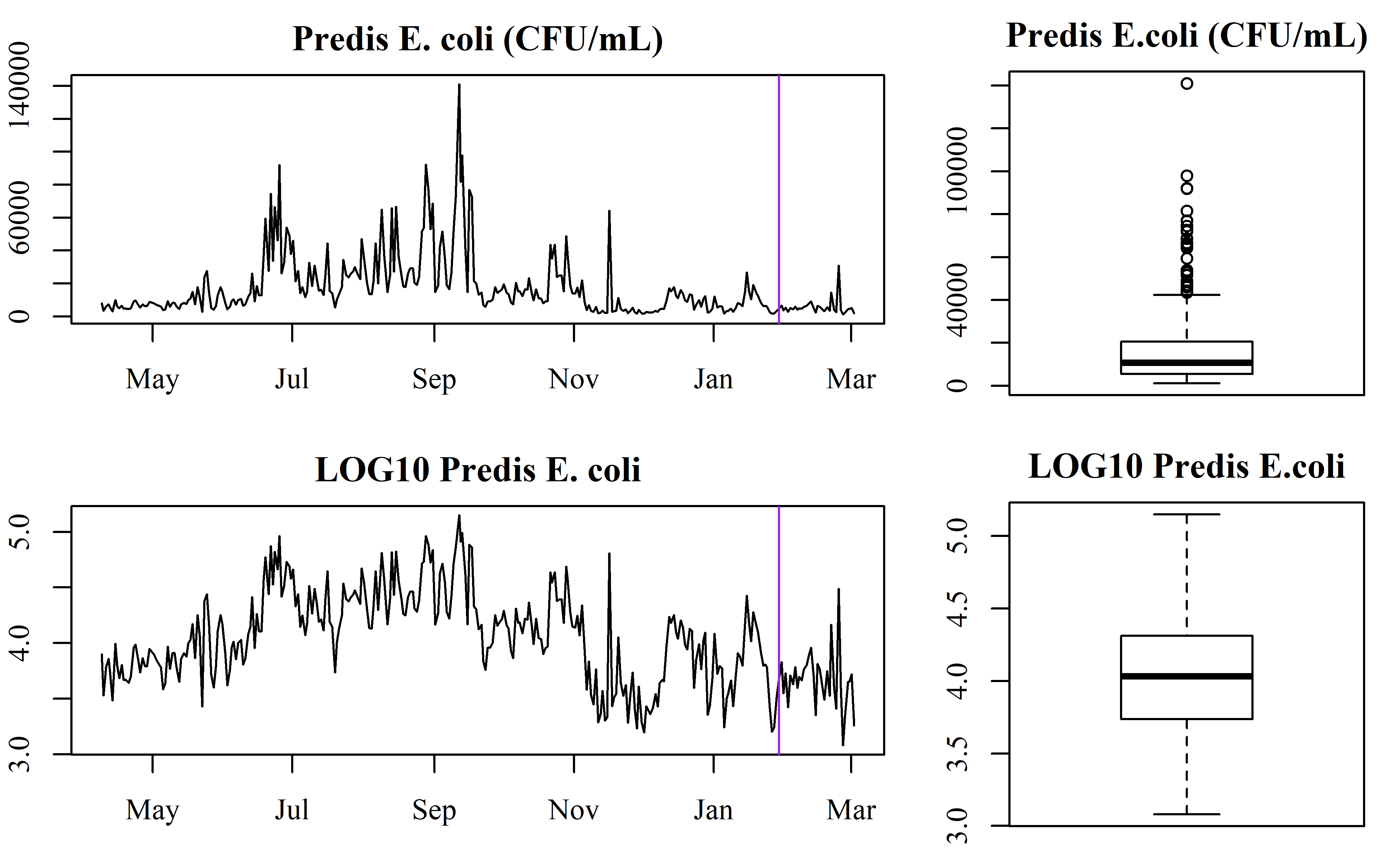
**Figure 1.** North predisinfection *E. coli* for training (before purple line) and testing (after purple line) period. A log10-transform approximates normally distributed training and testing data.



**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

| **Sorth Variable** | **Collection Method** | **Number of Observations** |
| --- | --- | --- |
| RWH South, Pri Pump Station, Pri Eff ALK | FC24 | 44 |
| RWH South, Pri Pump Station, Pri Eff BOD | FC24 | 304 |
| RWH South, Pri Pump Station, Pri Eff NH3A | FC24 | 314 |
| RWH South, Pri Pump Station, Pri Eff NO5 | FC24 | 128 |
| RWH South, Pri Pump Station, Pri Eff OP | FC24 | 40 |
| RWH South, Pri Pump Station, Pri Eff PW | FC24 | 131 |
| RWH South, Pri Pump Station, Pri Eff TIN | FC24 | 119 |
| RWH South, Pri Pump Station, Pri Eff TKNH | FC24 | 129 |
| RWH South, Pri Pump Station, Pri Eff TN | FC24 | 118 |
| RWH South, Pri Pump Station, Pri Eff TSS | FC24 | 312 |
| RWH South, Disinfection, Eff ALK | FC24 | 141 |
| RWH South, Disinfection, Eff BOD | FC24 | 175 |
| RWH South, Disinfection, Eff NH3A | FC24 | 323 |
| RWH South, Disinfection, Eff NO5 | FC24 | 134 |
| RWH South, Disinfection, Eff OP | FC24 | 39 |
| RWH South, Disinfection, Eff PW | FC24 | 133 |
| RWH South, Disinfection, Eff TIN | FC24 | 127 |
| RWH South, Disinfection, Eff TKNH | FC24 | 134 |
| RWH South, Disinfection, Eff TN | FC24 | 127 |
| RWH South, Disinfection, Eff TSS | FC24 | 180 |
| AB NO 2 ZONE 5 DO CTRL AC S123A | Online | 328 |
| AB NO 2 ZONE 6 DO CTRL AC S123B | Online | 328 |
| AB NO 2 ZONE 7 DO CTRL AC S123C | Online | 328 |
| AB NO 2 ZONE 8 DO CTRL AC S123D | Online | 328 |
| AB NO 6 ZONE 5 DO CTRL AC S163A | Online | 328 |
| AB NO 6 ZONE 6 DO CTRL AC S163B | Online | 328 |
| AB NO 6 ZONE 7 DO CTRL AC S163C | Online | 328 |
| AB NO 6 ZONE 8 DO CTRL AC S163D | Online | 328 |
| AI S568A DMX1 NH3 Analyzer | Online | 328 |
| AI S568D DMX2 NO3 Analyzer | Online | 328 |
| AI S578A DMX1 NH3 Analyzer | Online | 328 |
| AI S578D DMX2 NO3 Analyzer | Online | 328 |
| ASRT ASRT S | Online | 327 |
| CBOD | FC24 | 167 |
| COD | FC24 | 78 |
| ECIDX | Grab | 328 |
| FC T621 | Online | 328 |
| NH3A | Grab | 0 |
| NO5 | Grab | 0 |
| OP | Grab | 0 |
| PAA South Plant Flow | Online | 328 |
| S CMPLX SEC EFF FLOW FI F4 | Online | 328 |
| SSEC AB 1 IN SERVICE HC S110A | Online | 328 |
| SSEC AB 2 IN SERVICE HC S120A | Online | 328 |
| SSEC AB 2 NO3 AI S122D | Online | 328 |
| SSEC AB 2 ZONE 8 TSS AI S125C | Online | 328 |
| SSEC AB 3 IN SERVICE HC S130A | Online | 328 |
| SSEC AB 4 IN SERVICE HC S140A | Online | 328 |
| SSEC AB 5 IN SERVICE HC S150A | Online | 328 |
| SSEC AB 6 IN SERVICE HC S160A | Online | 328 |
| SSEC AB 6 ZONE 8 TSS AI S165C | Online | 328 |
| SSEC AB6 Z4 NH3 AI S163A 4A | Online | 328 |
| SSEC AB6 Z8 NH3 AI S163D 8A | Online | 328 |
| SSEC Ammonia Control PV | Online | 328 |
| SSEC CaRRB 2 SWAS TSS ANALYZER AI S75C | Online | 328 |
| SSEC CaRRB 2B NH3 AI S77A | Online | 328 |
| SSEC CaRRB 3 SWAS TSS ANALYZER AI S85C | Online | 328 |
| SSEC CaRRB Basin 2B NO3 AI S77A | Online | 328 |
| SSEC PEPS NH3 Analyzer AI S50A | Online | 328 |
| SSEC PEPS TSS Analyzer AI S50C | Online | 328 |
| SSEC RAS To SEC Influent Ratio | Online | 327 |
| SSEC Total CaRRB Flow to ABasins FY S102 | Online | 328 |
| SSEC Total PE Flow to ABasins FY S100 | Online | 328 |
| SSEC Total RAS Flow to ABasins FY S101 | Online | 328 |
| SSEC Total RAS Flow To CarrB FI S391 | Online | 328 |
| TI R3003 | Online | 327 |
| TOC | FC24 | 25 |
| TSS | Grab | 0 |
| TSSM | Grab | 0 |

**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

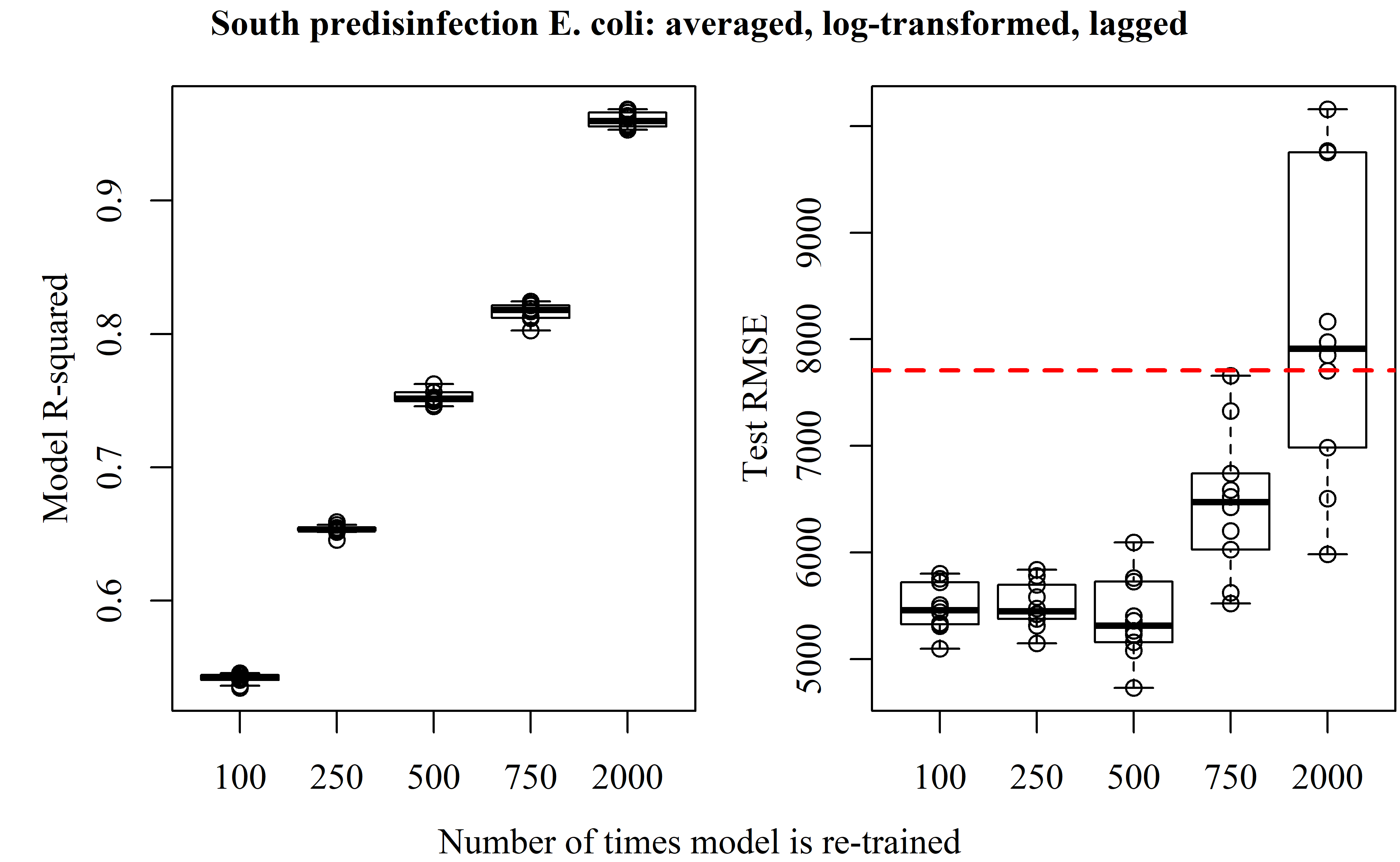
## *E. coli* prediction

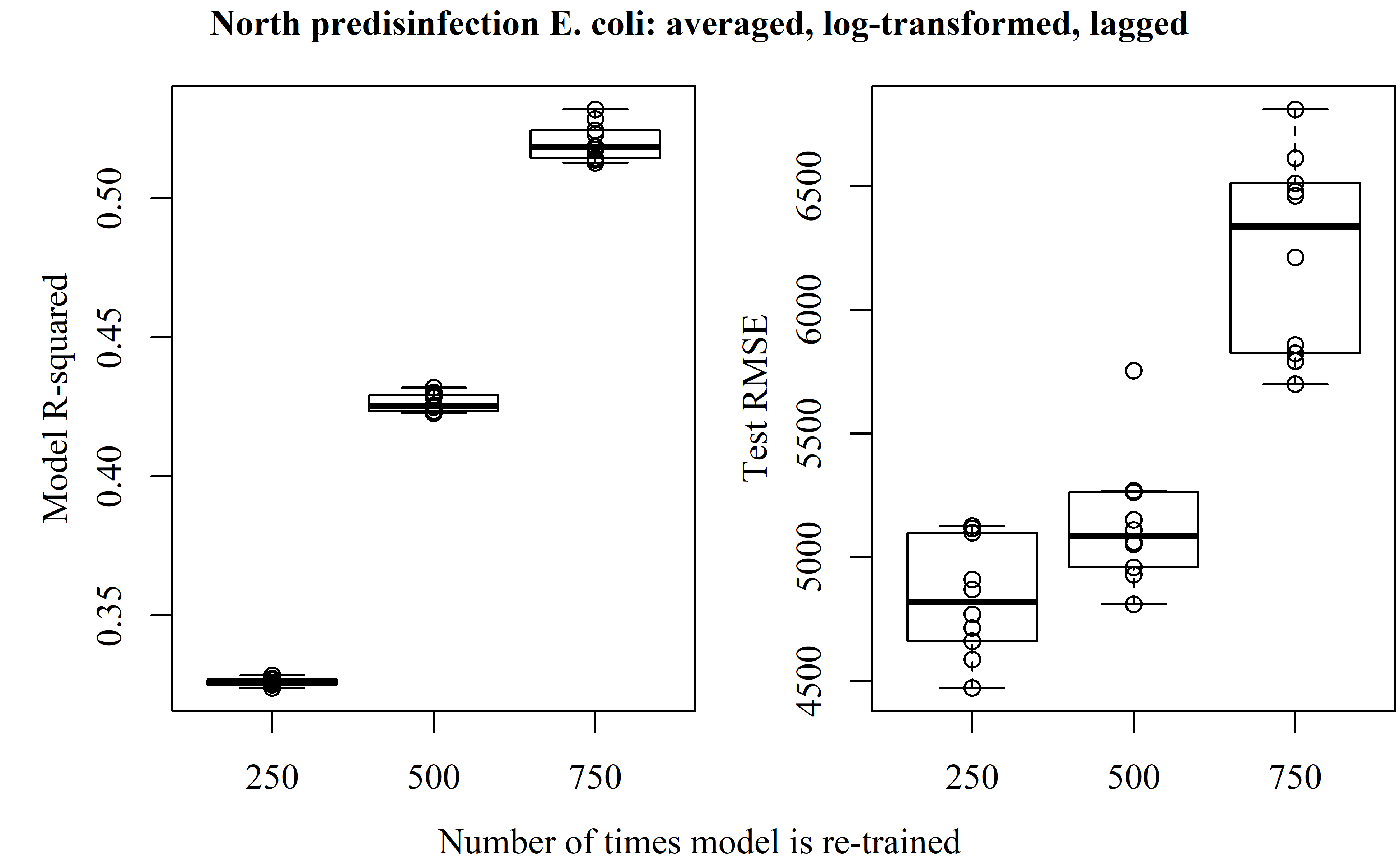
### Predisinfection *E. coli*

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.

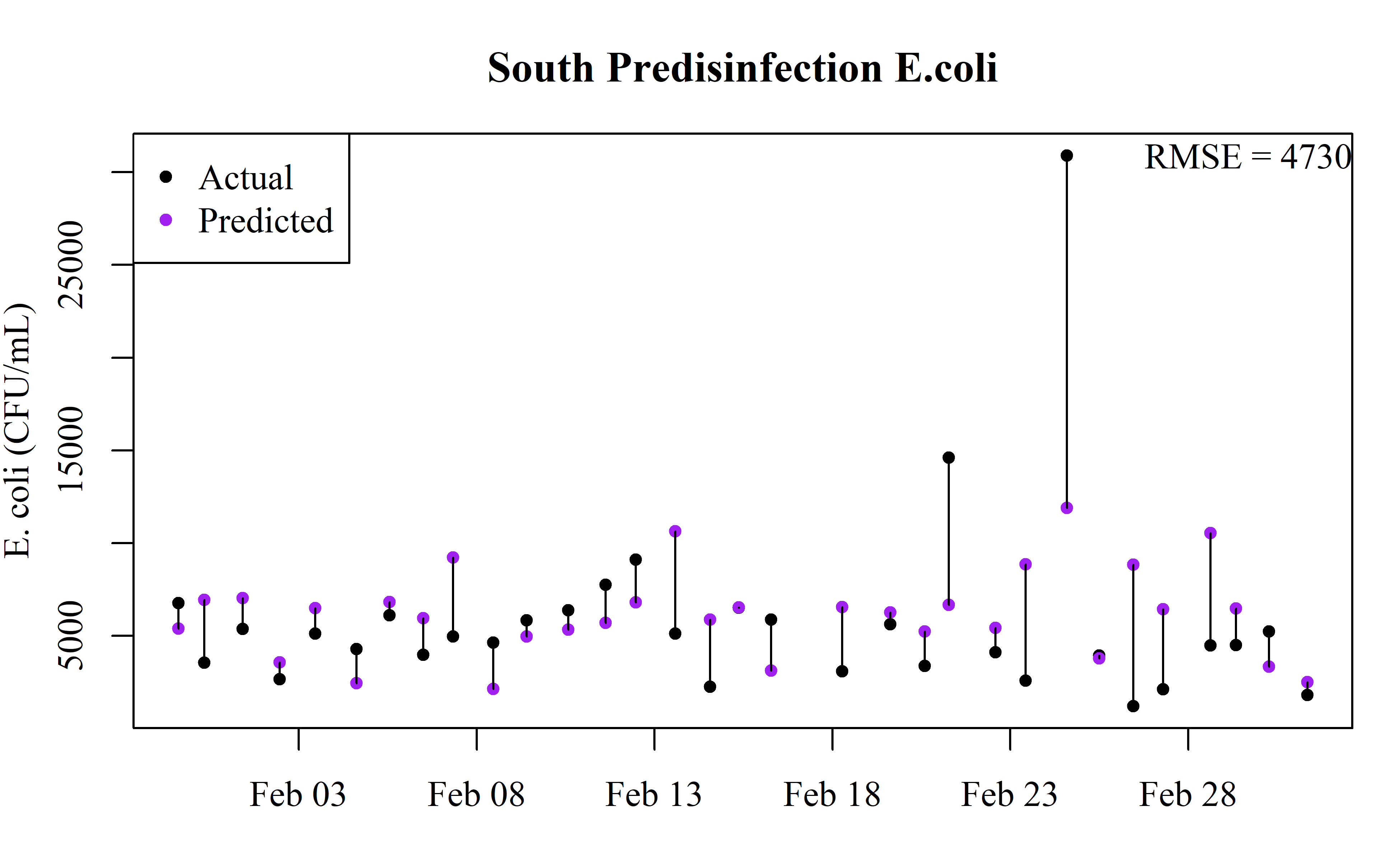
A common problem in machine learning is *overfitting*. Overfitting occurs when the model-building method continues to adjust model parameters beyond the general trends and instead fits to error within the training dataset. This causes a deceptively high training R2, as the testing RMSE is also very high. 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.

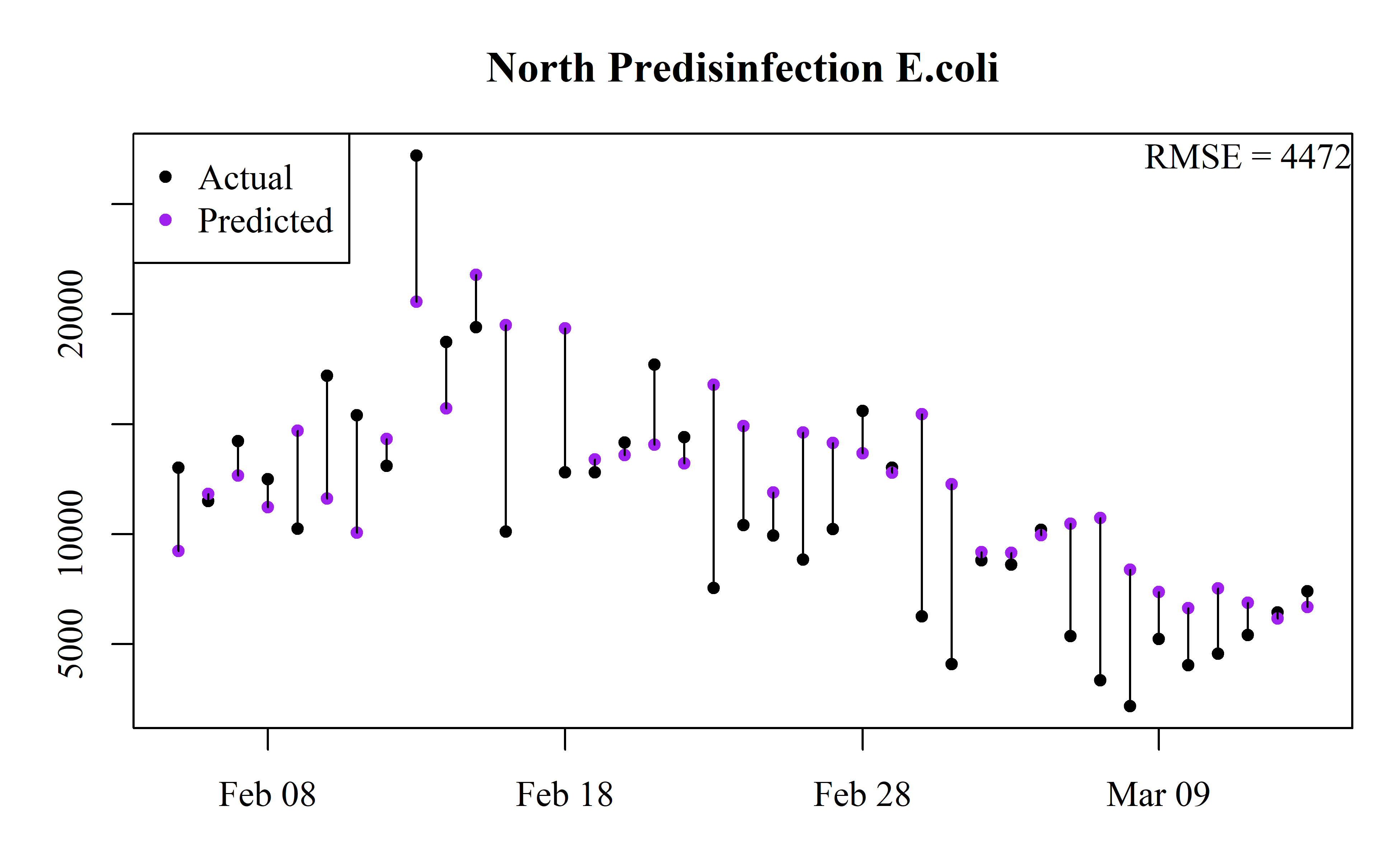
**Figure 3.** South predisinfection *E. coli* is predicted using a simple RNN, using a various number of epoch (e.g., “re-train” parameter estimations). The red line is the *persistence* RMSE, which is the RMSE calculated from the previously known *E. coli* value. Compared to the model predicted RMSE (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

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>.

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>.