**Introduction**

An extensive preliminary analysis of data associated with the peracetic acid (PAA) disinfection system was conducted, including methods such as linear regression, generalized linear models, random forest, support vector machines, generalized additive models, and principal component analysis (PCA). It was found that most data are not normally distributed, with no or weak linear and nonlinear relationships between variables. The variation in pre-disinfection *E.coli* and PAA demand (as captured by PCA) is independent of most process variables, but this could also be a consequence of large amounts of variation within the dataset. Artificial neural networks (NN) were selected in an attempt to account for the extensive nonlinear variation that is not captured by conventional models. Here, NN are used to predict PAA demand (difference between initial dose and PAA concentration approximately 1-min after dosing) and PAA decay (difference between PAA 1-min concentration and PAA concentration halfway through the disinfection basin).

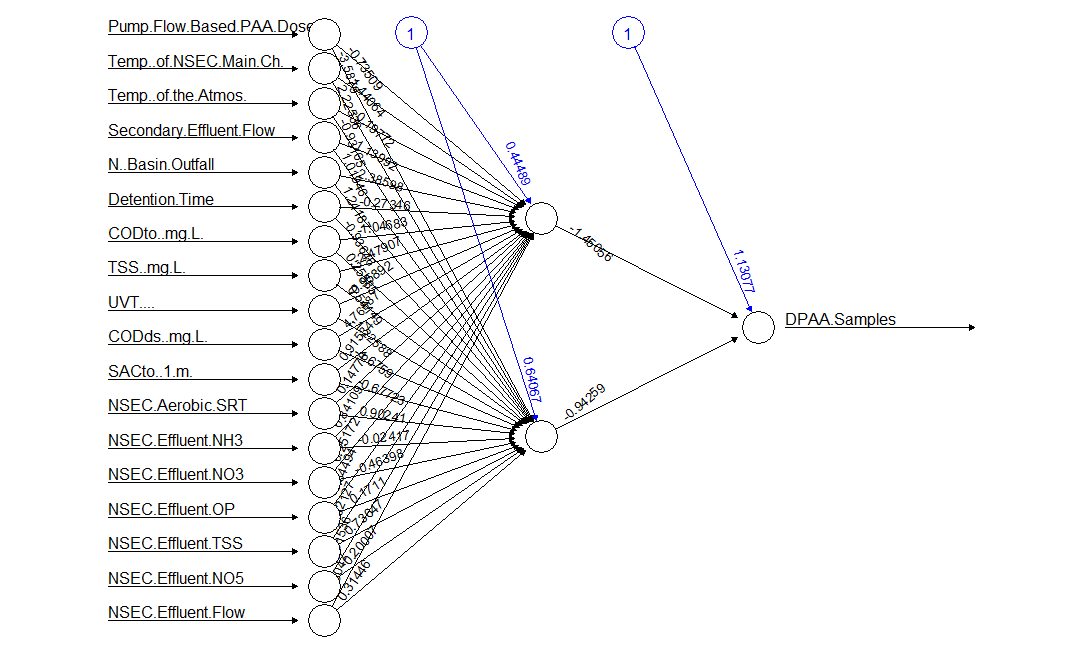
**Neural network model to predict PAA demand and decay**

A mixture of online and grab data was used to predict the total PAA demand and decay. To predict the change in PAA concentration throughout the North disinfection basin (from initial dosing to 1/2 basin), the following datasets were merged:

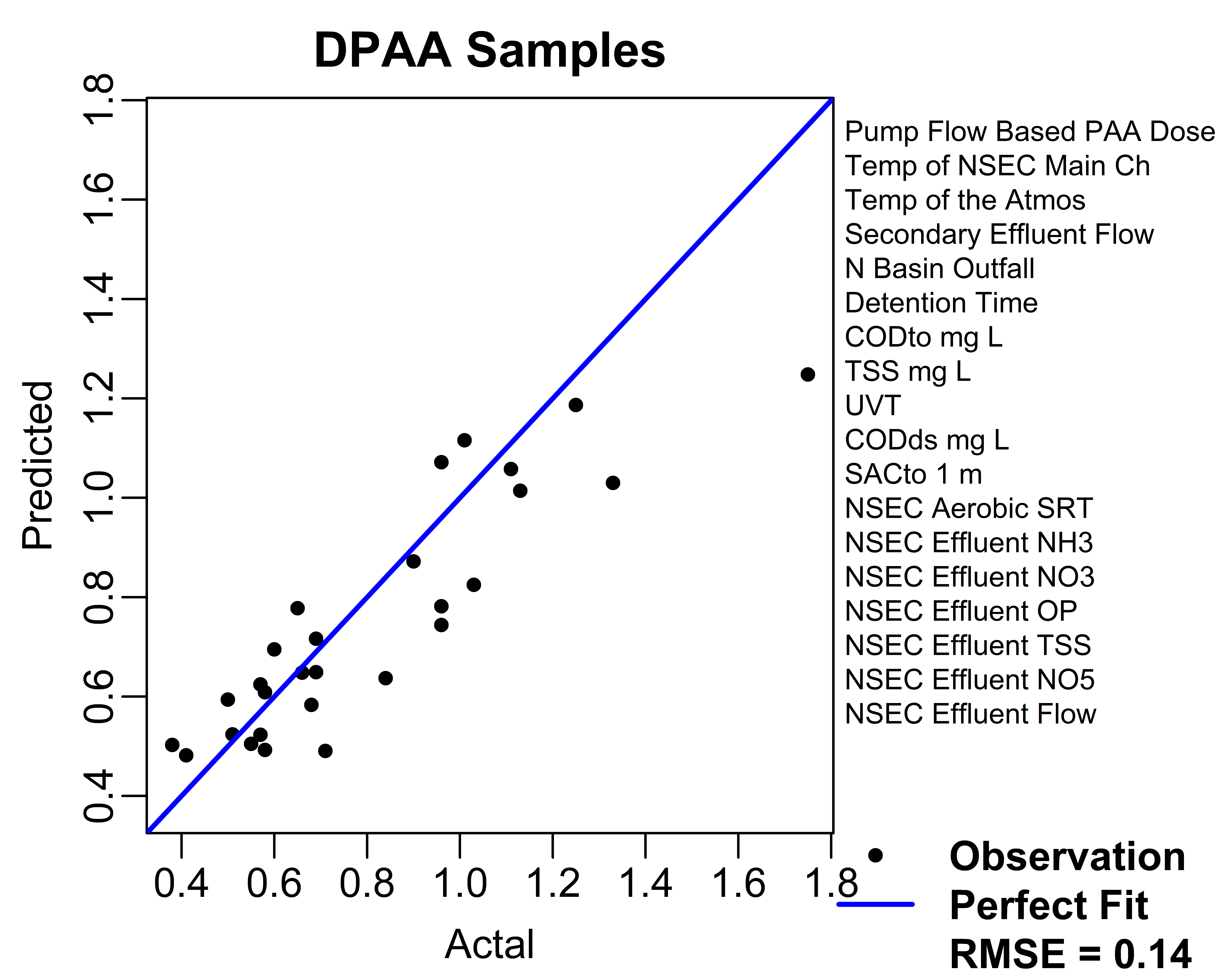
* 111 observations of PAA grab samples collected regularly between Oct 2-15, 2018
* Online data from a CarboVis analyzer upstream of PAA dosing
* Online data from the north secondary effluent

*PAA decay*

A randomly selected 80% of the data was used to train the NN model and the remaining 20% were used to validate the model. NN structure (**Figure 1**) was selected by minimizing the root-mean-squared-error (RMSE, standard deviation of prediction errors) assuming one hidden layer. *The best NN model (****Figure 2****) is able to predict PAA decay within 0.1 mg/L* when the decay is 1.2 mg/L. To improve the fit of the model between 1.2 mg/L and 1.4 mg/L, more observations within this decay range are needed.

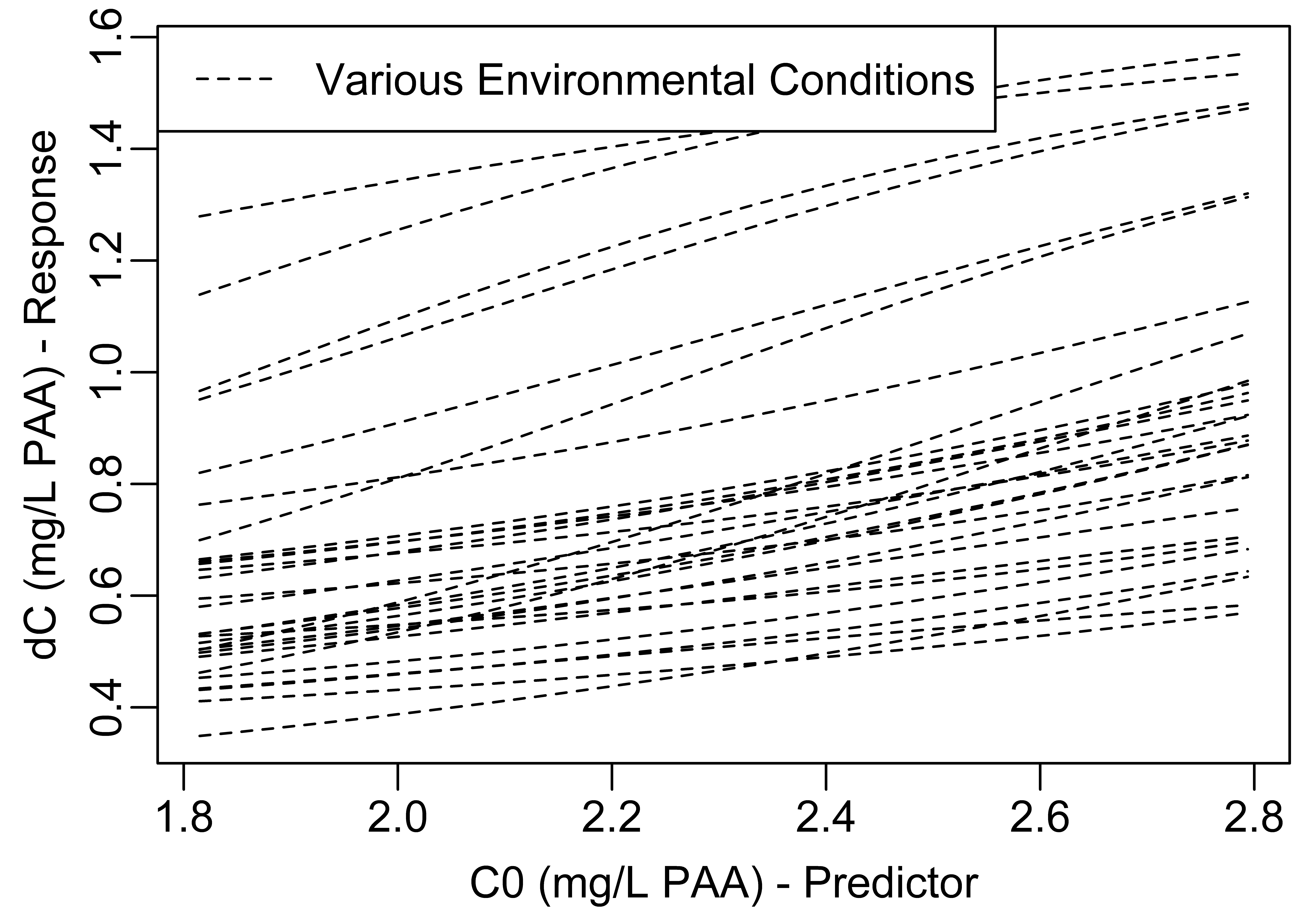


**Figure 1.** Example of the optimum NN structure used to calculate change in PAA concentration throughout the disinfection basin (“DPAA”). NN model was optimized by changing number of hidden nodes to minimize the RMSE; 2 are used in this example. At each node (circle), an activation function (sigmoid) transforms the weighted values (weights in black text) of each input plus a bias term (blue text). The weights and biases are determined by resilient back-propagation with weight backtracking.



**Figure 2.** Neural network model prediction of PAA decay (“DPAA” or Delta PAA) between the 1-minute PAA grab sample and the ½ basin PAA grab sample. Black dots are test data that include DPAA and the process variables listed to the right of the plot for each observation (i.e., point). RMSE is the root-mean-squared-error between the predicted and actual DPAA values and was used to identify the optimum NN model. The blue line is representative of a perfect model fit (i.e., predicted = actual).

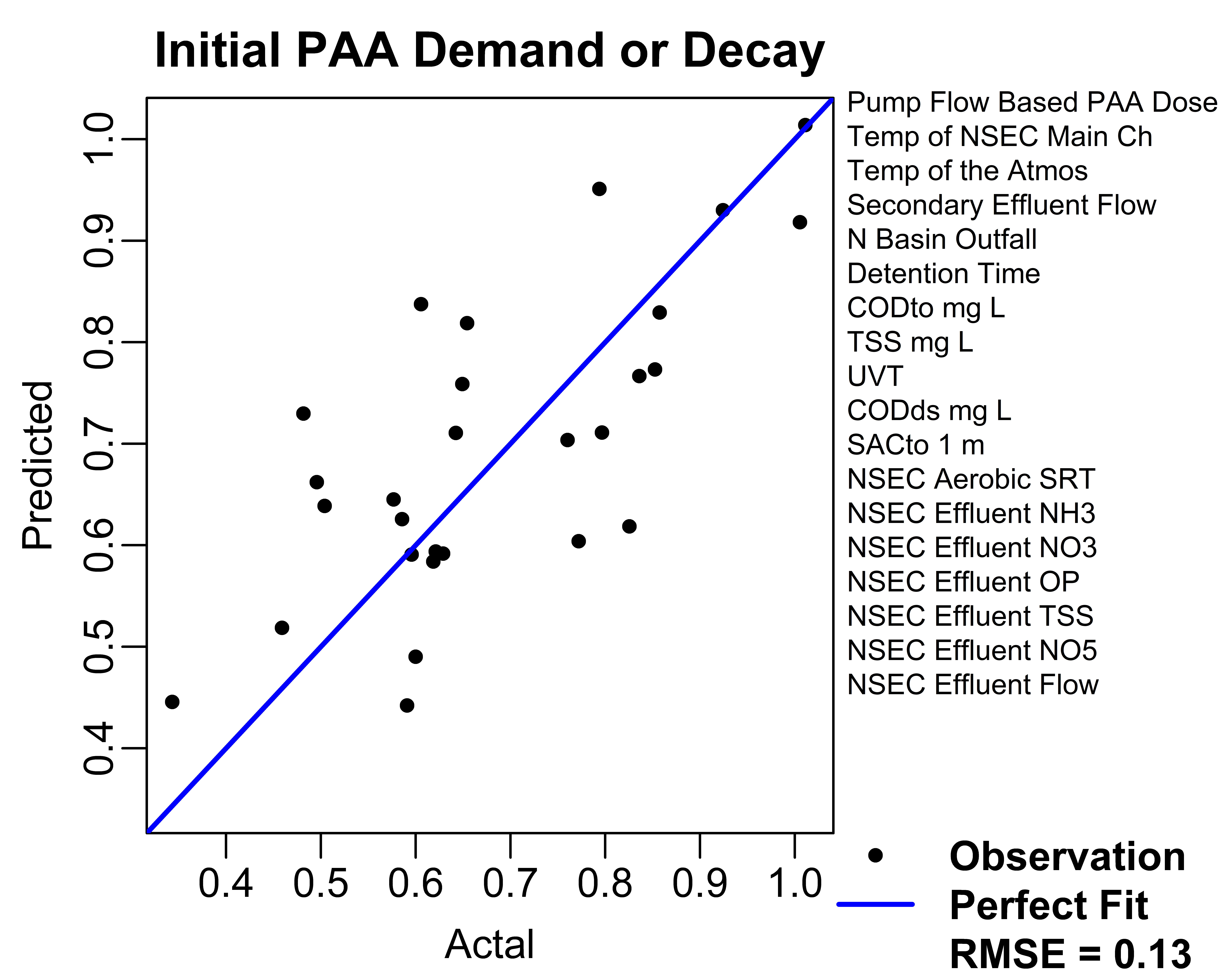
The impact of input variables to the NN model should not be underestimated. In **Figure 3**, PAA dosing flowrate was varied for all test cases (i.e., test observations from **Figure 2**). While the rate of PAA decay appears to vary linearly with PAA dose, it is dependent on additional water quality variables and only valid for a small range of initial PAA dosing concentrations.



**Figure 3.** For all test observations (i.e., environmental conditions) initial PAA dose was varied (x-axis) and PAA decay was predicted (y-axis) using the NN model from Figures 1 and 2. Plot demonstrates the impact of other water quality parameters considered in the NN model beyond initial PAA dosing rate that could impact the calculation of decay rate.

*PAA demand*

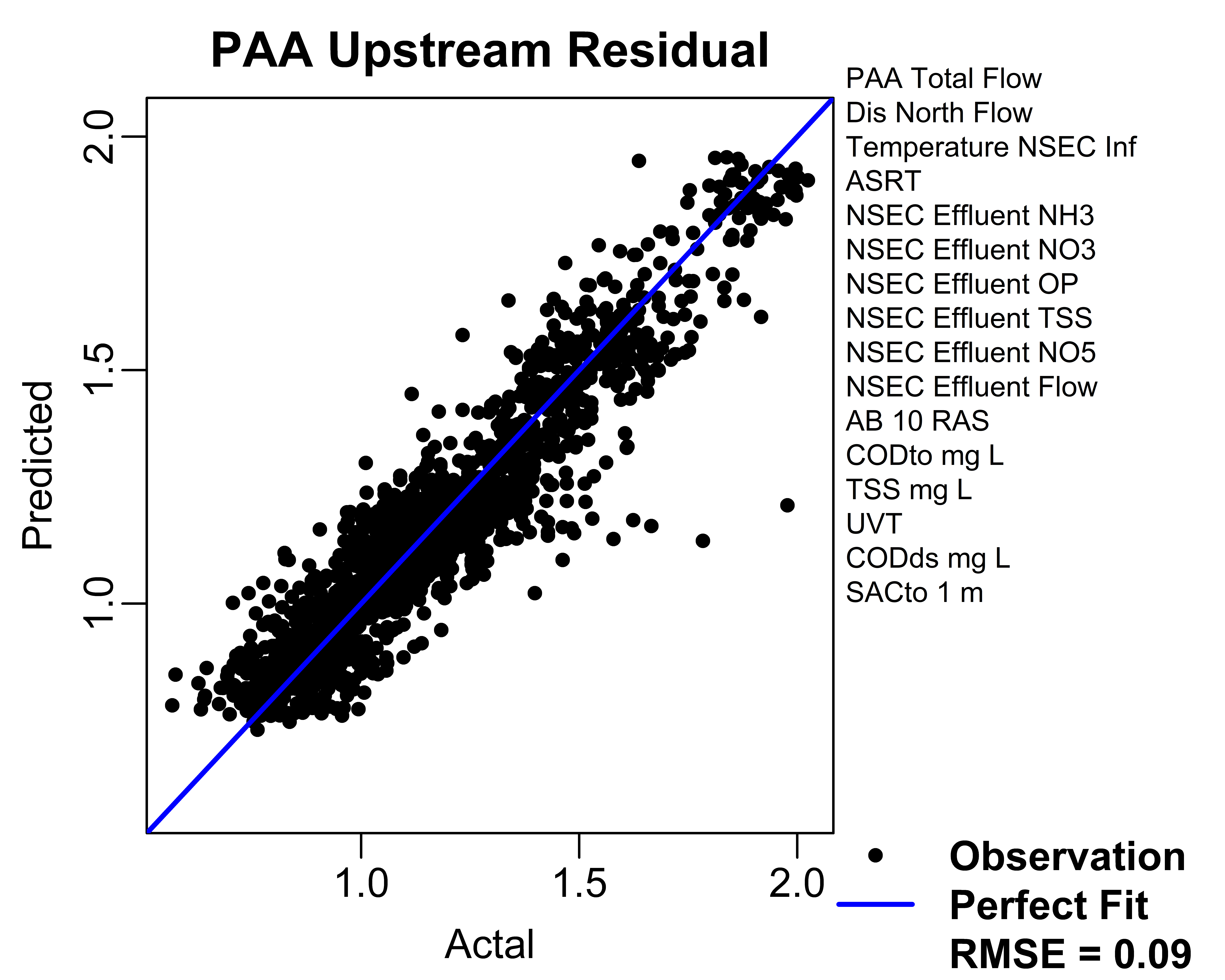
Following the same procedure performed for predicating PAA decay, instantaneous PAA demand was predicted. *The best NN model (****Figure 4****) is able to predict PAA demand within 0.2 mg/L*. While the dispersion appears to be greater in the prediction of PAA demand as opposed to PAA decay, this is a consequences of the range—the RMSE are approximately equal.



**Figure 4.** Neural network model prediction of the change in PAA concentration between the flow-based PAA dose and the 1-minute PAA grab sample (i.e., instantaneous demand or *D*). Black dots are test data that include initial PAA demand and the process variables listed to the right of the plot for each observation (i.e., point). RMSE is the root-mean-squared-error between the predicted and actual PAA demand and was used to identify the optimum NN model. The blue line is representative of a perfect model fit (i.e., predicted = actual).

**Neural network models – Grab vs Online PAA**

To compare how initial PAA demand is predicted, grab data from October 2018 (**Figure 4**) and online data from December 2018 (**Figure 5**) was used to build predictive NN models. The range of online data is limited (> 0.5 mg/L PAA), but functions well for prediction of immediate PAA demand. *Predicting online PAA at the 1-minute sample point is slightly more accurate using online PAA data as opposed to grab samples* (Online RMSE 0.9 vs Grab RMSE 0.13)



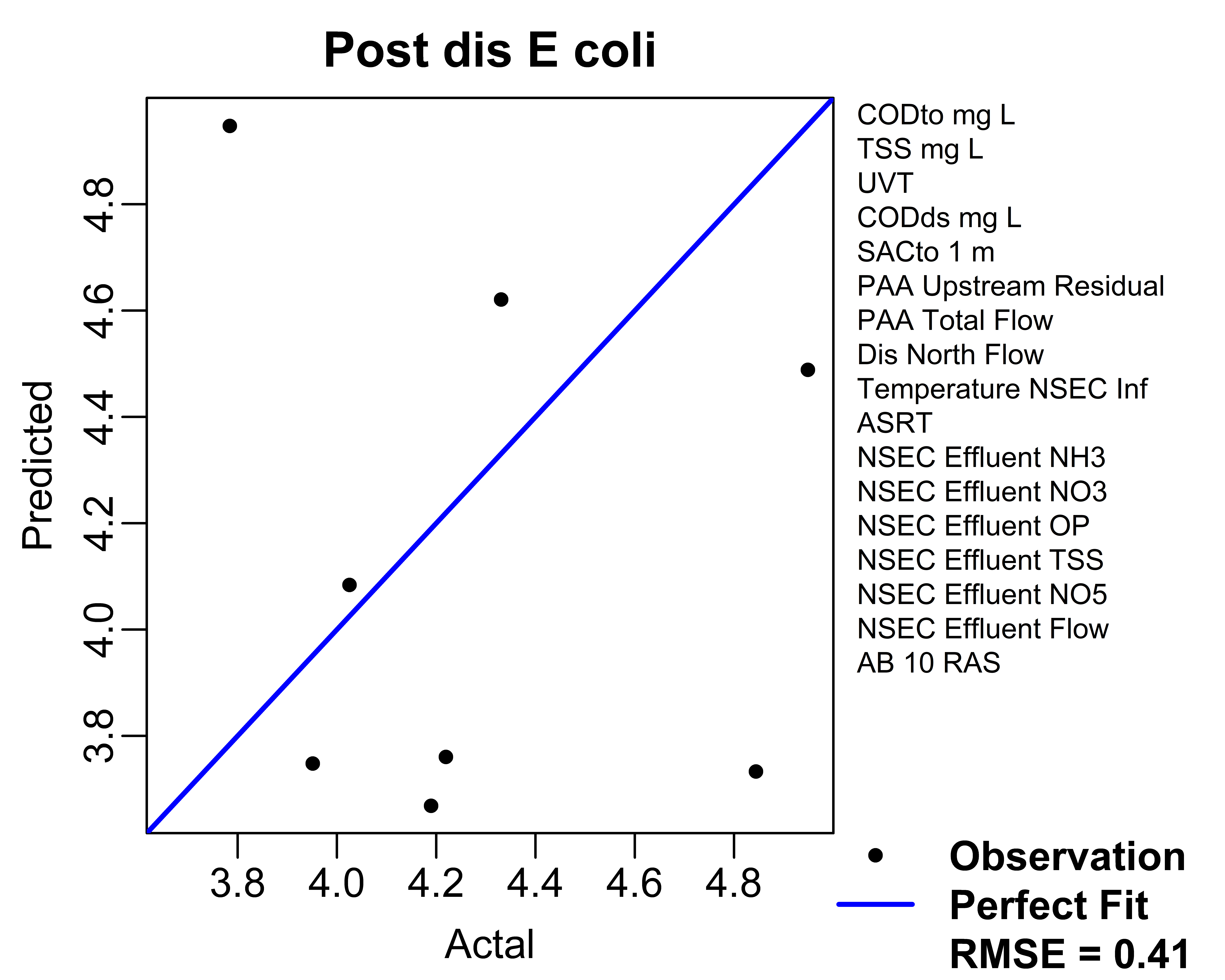
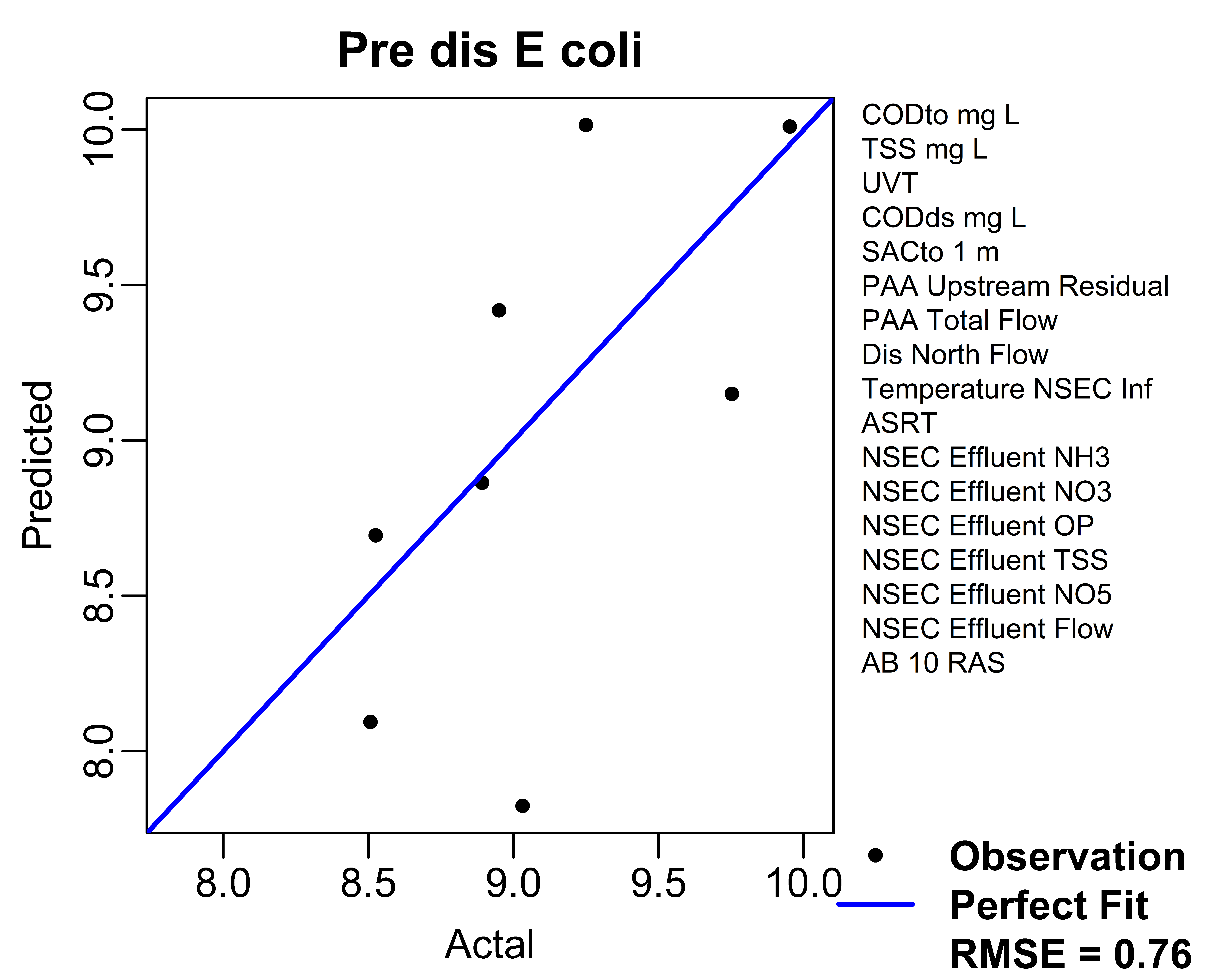
**Figure 5.** Neural network model prediction of the 1-minute PAA concentration from the Chemscan online PAA analyzer. Black dots are test observations and the blue line is representative of a perfect model fit.

**Neural network model to predict *E. coli***

The following datasets were used to predict pre and post disinfection *E. coli* in the disinfection basin:

* Online PAA data (Chemscan)
* Online water quality data (CarboVis and nutrient sensors in North Secondary)
* Daily *E. coli* grab samples

Due to limited overlap with regards to the date ranges of each dataset, there were few observations to train (31) and test (8) the model, lending to a substantial amount of error (**Figure 6**). The major limitation was the availability of water quality data from the CarboVis online analyzer since the Chemscan has been online.

****

**Figure 6.** NN model prediction for pre-disinfection (left) and post-disinfection (right) *E. coli*. The standard deviation of pre-disinfection *E. coli* (RMSE = 0.76) is approximately twice the post-disinfection *E. coli* (RMSE = 0.41), indicating there are variables impacting initial *E. coli* concentrations that are missing from the NN model.

**Main takeaways**

* PAA concentrations along the disinfection basin at the 1-minute and ½ basin sample points can be accurately predicted using artificial neural networks and online sensor data (Figures 2,4,5)
* *E. coli* concentrations cannot yet be accurately predicted using artificial neural networks, most likely due to a limited dataset (Figure 6)
* First-principal model parameters, *D* and *k*, most likely vary nonlinearly given the wide range of operating and environmental conditions of the PAA system (Figure 3)

**Remaining questions**

* Impact of each variable to the neural network?
  + What happens if a sensor is taken offline for an extended period of time?
  + What happens if a sensor is taken offline for a short period of time for calibration?
  + Does the model need to be retrained after every calibration?
* Can ICT be directly predicted using neural networks?

**To do**

* CarboVis controller needs to be connected to the online database and assign PI tags
* Re-evaluate neural network *E.coli* predictions with more CarboVis data, solar UV data
* Use NN model predictions for PAA concentrations to calculate *k* and *D* for first-principal model
* Generate ICT curves from NN predictions