Long-Term Vulnerability Assessment and Adaptation Planning for the San Francisco Public Utilities Commission Water Enterprise

Technical Report 5: Raw Water Quality Model

Prepared for:

San Francisco Public Utilities Commission

Prepared by:

Hydrosystems Research Group, University of Massachusetts Amherst

May 31, 2021

Summary

This Technical Report describes the development of data-driven models to simulate the raw water quality of the Hetch Hetchy Regional Water System (HHRWS) to be used in the Long-Term Vulnerability Assessment (LTVA) and the Adaptation Planning for the San Francisco Public Utilities Commission Water Enterprise. Raw water quality models aim at predicting turbidity at the O'Shaughnessy Dam; turbidity and total organic carbon (TOC) at the Tesla Portal in HHRWS. Based on the performances of several prediction methods, the Composite Quantile Regression Neural Network (CQRNN) has been selected for the LTVA. CQRNN's inputs include various hydrometeorological variables (e.g., inflows, air temperature, and water temperature), system state variables (e.g., reservoir storage, spill and release, flows through relevant pipelines), and variables that summarize the system memory (e.g., flows and dry days for previous and current water years, antecedent dry days). Historical observations show that levels of raw water quality had exceeded normal operating targets for turbidity at the Tesla Portal in 1997, which shut down the San Joaquin Pipeline for several days. To simulate the water system response to levels of raw water quality exceeding normal operating targets, it would require a feedback loop between the San Francisco System Model (HRG TR4 2021) used to simulate the operation of the HHRWS, and the considered CQRNN, which is not the option that has been chosen here. Instead, results of CQRNN have been used to develop water quality narratives as described in the LTVA report (HRG LTVA 2021). Furthermore, developed CQRNN models have revealed the system vulnerability regarding raw water quality via a climate stress test. This report describes the goal, the calibration, and the validation of the CQRNN approach. Results are obtained for turbidity at the O'Shaughnessy Dam, turbidity, and TOC at the Tesla water treatment plant. The performance of CQRNN for raw water quality prediction in HHRWS is compared with four state-of-the-art approaches commonly used in the literature (Quantile Regression, Linear Regression, Multivariate Adaptive Regression Spline, and K-Nearest Neighbors). The report ends with a conclusion that summarizes the model development.

Table of Contents

List of Abbreviations

ADD	Antecedent dry days
AF/af	Acre-foot
CF/cf	Cubic feet
CFD/cfd	Cubic feet per day
CFS/cfs	Cubic feet per second
HH	Hetch Hetchy
HHRWS	Hetch Hetchy Regional Water System
KNN	k-Nearest Neighbors
LR	Linear regression
LTVA	Long-Term Vulnerability Assessment and Adaptation Planning for the SFPUC Water Enterprise
MARS	Multivariate Adaptive Regression Spline
mg/L	Milligram per liter
MGD/mgd	Millions of gallons per day
MSE	Mean squared error
NTU	Nephelometric Turbidity Unit
OSH	O'Shaughnessy Dam
PCT	Percentile
CQRNN	Composite Quantile Regression Neural Network
SFPUC	San Francisco Public Utilities Commission
SFWSM	San Francisco Water System Model
SJPL	San Joaquin Pipeline
TOC	Total organic carbon
Turb	Turbidity

List of Figures

List of Tables

1 Introduction

1.1 Objective

Climate change and other changing conditions may jeopardize the Hetch Hetchy Regional Water System (HHRWS) to meet the San Francisco Public Utilities Commission (SFPUC) in the future. To help address this concern, SFPUC collaborated with the University of Massachusetts Amherst (UMASS Amherst) to help identify critical vulnerabilities of the HHRWS to long-term changes in climate and other conditions under a project called the Long-Term Vulnerability Assessment and Adaptation Planning for the SFPUC Water Enterprise (LTVA). The objective of the LVTA is, in part, to "design and execute an exhaustive vulnerability assessment that provides a comprehensive understanding of the expected water system performance relative to goals and expectations of the system." The general approach to achieve this objective, as outlined in the LTVA Detailed Analytical Plan, is to "develop a suite of interconnected computer models and supporting analytical modules representing important processes involved in the long-term planning of HHRWS and to then use them along with a scenario discovery approach to quantitatively assess system vulnerability." This report focuses on developing a machine learning approach to predict raw water quality (i.e., turbidity and TOC) at the O'Shaughnessy Dam and the Tesla Portal.

1.2 Context and scope of the report

SFPUC conveys unfiltered water from the snowmelt-dominated Tuolumne River to the SFPUC service area. The Hetch Hetchy reservoir is the key infrastructure of the HHRWS, controlling the water flow to the Tesla water treatment facility via the San Joaquin Pipeline (SJPL). Hetch Hetchy water must meet all federal and state filtration avoidance criteria at the Tesla Portal to ensure no additional water treatment before being supplied to customers. The water enters the Tesla treatment facility to be disinfected using UV light and chlorine disinfection for the Hetch Hetchy supply. If turbidity, total organic carbon, and fecal coliform levels comply with the filtration avoidance criteria, water can be delivered to customers; otherwise, water must be filtered at the Sunol valley water treatment plant.

A major turbidity event happened in January 1997 and is an example of a violation of the filtration avoidance criteria (Figure 1). A major rain-on-snow event caused the highest turbidity peak recorded at the Tesla portal (i.e., 6.77 NTU or 6.77 Nephelometric Turbidity Units). Eventually, it led to a shutdown of the SJPL. The high turbidity event occurred on 01/03/1997, caused by the big storm on 01/02/1997 (Figure 1). Such a high turbidity value is rare at the Tesla portal. The second and third highest turbidity peaks recorded at the Tesla Portal are 2.65 NTU (2000-11-22) and 2.5 NTU (1996-11-24). The recurrence interval of the 1997 turbidity event has been estimated to 67 years (see the detail of the frequency analysis in the appendix); while the recurrence interval of the second and third highest turbidity peaks ever recorded at the Tesla Portal

is 12 years (2.65 NTU on 2000-11-22) and 11 years (2.5 NTU on 1996-11-24). We find that the 1997 turbidity event is rare, whose recurrence interval is six times greater than the recurrence interval of the second-highest turbidity.

The 1997 turbidity event summarizes a severe concern of SFPUC regarding climate change. Would events like the 1997 event become more frequent under climate change? What would be the consequences for water delivery in the Bay area? Would climate change emphasize other processes that could ultimately lead to degradation of the Hetch Hetchy water quality?

SFPUC needs to advance understanding of the vulnerability of the HHRWS regarding potential shifts in climate and other regulations that could lead to the alteration of water quality from the Hetch Hetchy reservoir. For this purpose, raw water quality models are developed to simulate the temporal dynamic of turbidity at the O'Shaughnessy Dam, turbidity, TOC at the Tesla Portal in the HHRWS, and their responses regarding potential changes in climate and other environmental regulations.

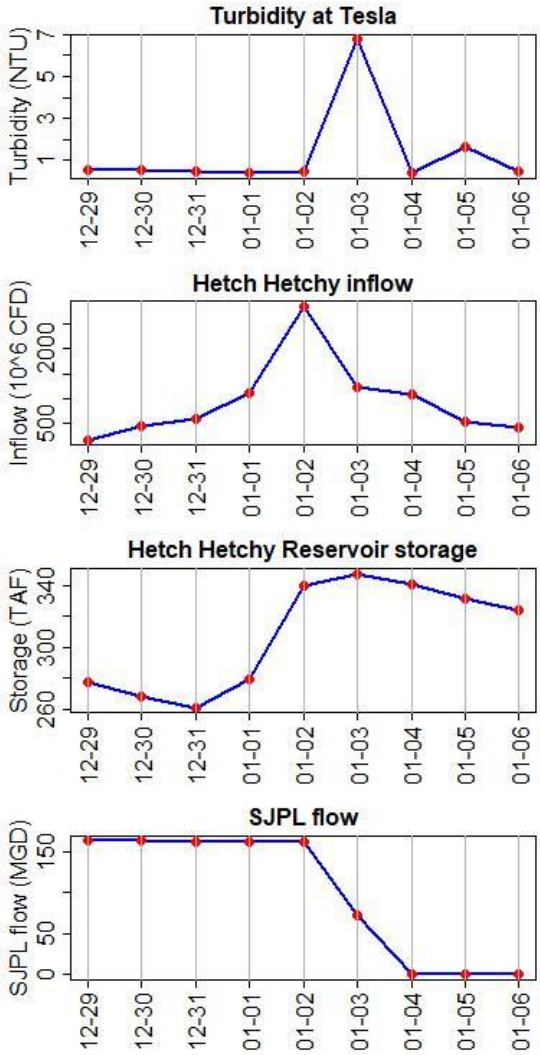

Figure 1 From top to bottom, turbidity at the Tesla Portal, the Hetch Hetchy inflow, the Hetch Hetchy Reservoir storage, and San Joaquin Pipeline flow from December 29, 1996, to January 6, 1997. The big storm on January 2, 1997 (second figure from the top) caused the highest turbidity on January 3, 1997 (top figure)

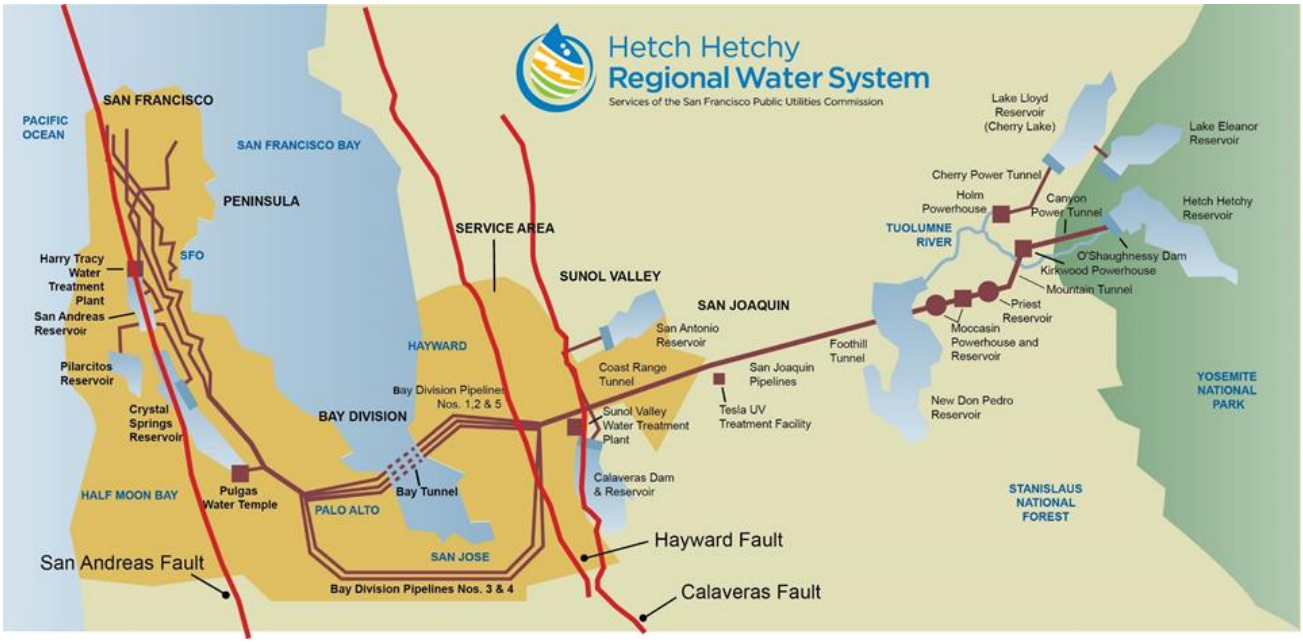

Figure 2 Hetch Hetchy Regional Water System owned and operated by SFPUC. Raw water quality models focus on the region between the Hetch Hetchy Reservoir and the Tesla Portal

The vulnerability assessment of the SFPUC system regarding the water quality is conducted following the decision scaling approach (Brown et al. 2012). The flow chart in Figure 3 illustrates this approach. This report describes the development and assessment of raw water quality models. The climate stress test results are included in the Vulnerability Assessment Final Report (HRG LTVA 2021).

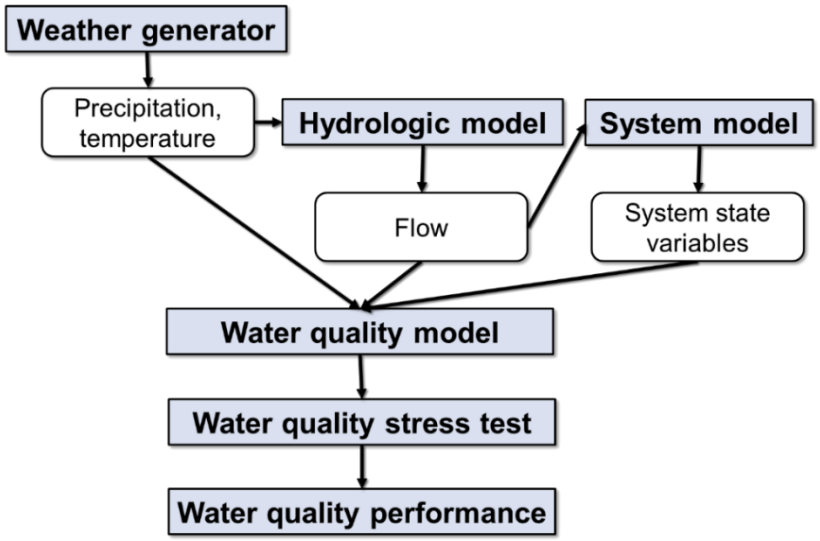

Figure 3 Decision Scaling Approach for the Water Quality Vulnerability Assessment

The UMass Amherst research group receives data of five indicators (turbidity, TOC, Escherichia coli or E.coli, fecal coliform, total coliform) at several locations (Priest Bypass, the Tesla Portal, Priest Reservoir, O'Shaughnessy Dam, Mountain Tunnel at Priest, Moccasin Gate Tower/Moccasin Reservoir, Kirkwood, West Portal). For E.coli, fecal coliform and total coliform, chemical and biological input data should be included in models. However, only physical data such as hydroclimatic and operational factors are available. Thus, fecal and total coliform indicators are not considered in the study. SFPUC agrees that turbidity and TOC are the main studied water quality indicators. Also, they confirm that the O'Shaughnessy Dam and the Tesla Portal are focused on this project. The model is not built for TOC at the O'Shaughnessy Dam because the UMass Amherst team does not receive the input data.

- **Turbidity** – Turbidity is a measure of the light scattered by suspended particles in the water and, as such, is a tracer of the concentration of suspended sediments. It is an indicator of water pollution.
- **Total Organic Carbon (TOC)** – TOC is the amount of carbon in organic compounds. It is a tracer of organic substances. Furthermore, TOC can combine with the disinfectant in RWS (free chlorine) to cause disinfection byproducts (DBPs).

To maintain water quality at acceptable levels, SFPUC needs to monitor water quality at different locations of the HHWRS and trigger adapted operations depending on the type of contamination, if required. For instance, if the turbidity and TOC levels of the Hetch Hetchy water exceed the filtration avoidance criteria, the water must be treated at the Sunol Valley Water Treatment Plant (SVWTP). Specific filtration avoidance thresholds are set for each indicator of water quality:

- **Turbidity** – The objective of SFPUC is to keep turbidity below 1 Nephelometric Turbidity Units (NTU). A threshold of 2 NTU is possible but may be detected by wholesale customers. The filtration avoidance threshold is 5 NTU (2 times in a 12-month rolling window).
- **Total Organic Carbon (TOC)** – TOC should not exceed a level of 2 mg/L, and DBP violations likely occur when TOC is above 3 mg/L.

1.3 Literature review

Delpla et al. (2009) reviewed the studies assessing the potential effects of climate change on surface water quality. They conclude that innovative predictive tools are required to enhance risk management and adaptation measure.

For example, Samal et al. (2013) studied the effect of climate change on the water turbidity in the Ashokan Reservoir (New York), one major reservoir for New York City's water supply system. The Ashokan Reservoir receives water from the Esopus Creek watershed, which originated from Slide Mountain. New York City's water supply system is located in a snow-dominated watershed, comparable to the upper part of the HHRWS. Across the upstream catchment of the Ashokan

reservoir, rain-on-snow events lead to high turbidity in the reservoir in Spring and Fall seasons, as shown in Figure 4. Figure 4 also illustrates the effect of climate change on turbidity in the Ashokan reservoir. During winter months (i.e., from December to March), the warmer temperature increases the fraction of liquid precipitation, leading to higher flows that may re-suspend and erode materials in the riverbed. It increases turbidity in the reservoir. The turbidity decreases from current to future climate conditions in the early Spring season (i.e., April and May). A reduced snowpack can be explained by warmer temperatures in Winter, which further decreases flow to the reservoir consequently reduces turbidity.

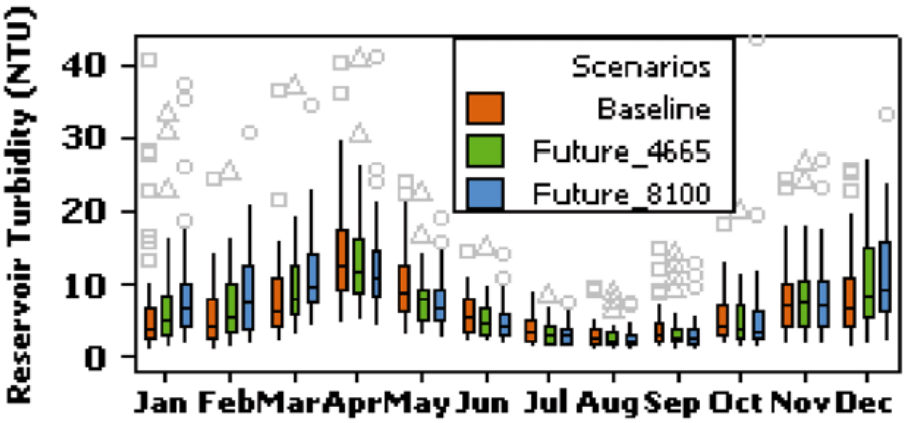

Figure 4 Average monthly reservoir turbidity under baseline and the simulated future period 2046-2065 (Future_4665) and 2081-2100 (Future_8100) (Samal et al. 2013a)

As illustrated in Figure 5, Weyhenmeyer and Karlsson (2009) showed that TOC levels increase nonlinearly with air temperature and runoff by using 3,123 lake water samples. This figure illustrates the effect of the warming temperature on TOC concentration for snow-dominated watersheds. As the average air temperature rises, the period during which the vegetation grows extends. The presence of flora increases the number of organic materials available to be flushed out during the high flow event, eventually leading to higher TOC values during these events.

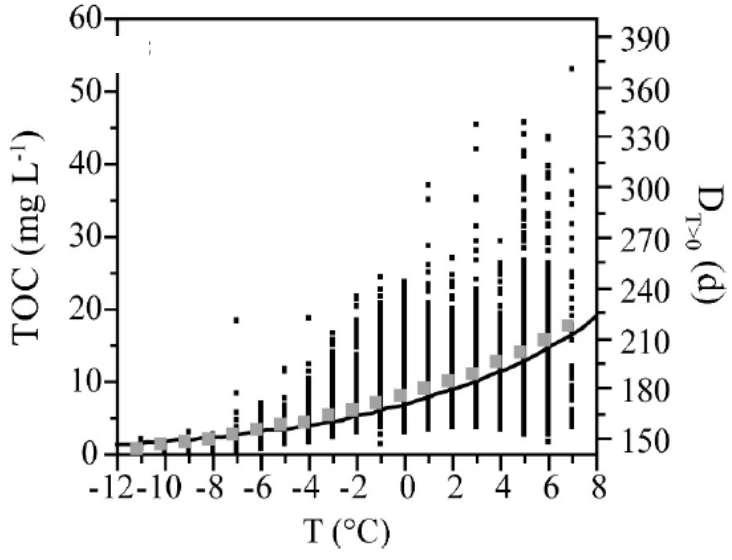

Figure 5 Total organic carbon (TOC) of 3,123 lake water samples along with lake-specific annual mean air temperatures (T). TOC increases nonlinearly with increasing temperatures, clarified by mean values per degree temperature (gray dots). The nonlinear increase corresponds to the nonlinear increase in the duration of the main growing and runoff season ($D_{T>0}$; black line) (Weyhenmeyer and Karlsson 2009)

1.4 Motivation for a new raw water quality prediction tool

SFPUC currently uses empirical models for simulating turbidity at O'Shaughnessy Dam and TOC at the Tesla Portal (Rob Clark 2017) (Table 1). However, SFPUC does not have the model for turbidity at the Tesla Portal. For turbidity at the O'Shaughnessy Dam, SFPUC uses a linear regression model at a weekly time step. For TOC, SFPUC uses a linear regression model for the annual scale. These models have several drawbacks:

1. Time step: Weekly and annual time steps do not allow daily water quality monitoring, which may be essential to adapt HHRWS operations if needed. For instance, the example of the turbidity event in 1997 highlights that significant changes in water quality can happen from one day to another, which motivates developing a modeling framework that uses a daily time step.
2. The recent literature has demonstrated that linear regression has a limited ability to capture nonlinear relations between water quality indicators and hydrometeorological (Futter and de Wit 2008; Weyhenmeyer and Karlsson 2009; Dhillon and Inamdar 2013; Mukundan et al. 2013).
3. Two recent linear regression models from SFPUC (weekly turbidity at O'Shaughnessy Dam, and annual TOC at the Tesla Portal) (Rob Clark 2017) lack the validation recommended for ensuring the robustness of the approach, which is paramount when intended to be used for a long-term vulnerability assessment.
4. SFPUC's model is not designed to predict high water quality events, i.e., peaks (Rob Clark 2017). Because extreme events are rare by definition, the 'weight' attributed to extreme events for estimating the regression parameters is a fortiori small compared with the ''mass'' of observed values between 20th and 80th percentiles. The Composite Quantile Regression Neural Network (CQRNN), developed for the need of the LTVA, can provide several quantiles for the water quality prediction. A simple regression model only gives a single estimate, usually by minimizing the mean cost function between observed and simulated values. By providing the distribution (i.e., several quantiles), we account for the uncertainty on the prediction.
5. SFPUC models only simulate turbidity at the O'Shaughnessy Dam and TOC at the Tesla Portal, while we have one additional model: turbidity at the Tesla Portal.
6. SFPUC models only consider hydrometeorological input, whereas the water quality might be affected by system operations described by state variables such as reservoir levels, pipeline flow, spill, release, etc. CQRNN enables us to link the water quality with both hydrometeorological predictors and system state variables.

The UMass Amherst research team has developed a new prediction tool for raw water quality to address the aforementioned limitations.

Table 1 compares CQRNN with recent water quality models at SFPUC and reveals the strengths of CQRNN. The main advantages of CQRNN are:

1. CQRNN predicts water quality indicators using both hydrometeorological conditions and system operations.
2. The developed CQRNN provides several quantiles for water quality prediction. For the LTVA, low (2.5^{th}, 25^{th}), median (50^{th}), and high (75^{th}, $97,5^{th}$) percentiles of distribution for the considered water quality indicators are modeled. It allows the description of the uncertainty around the median prediction (50^{th}) using a 95% confidence interval.
3. By analyzing the CQRNN model parameters, the contributions from each predictor to the prediction can be identified.

Table 1 Compare the SFPUC's recent water quality model and the Composite Quantile Regression Neural Network

Characteristic	SFPUC model for weekly turbidity at the O'Shaughnessy Dam	SFPUC model for annual TOC at the Tesla Portal	The proposed model in this study
Method	Linear regression	Linear regression	Composite Quantile Regression Neural Network (statistical model, i.e., quantile regression, combined with machine learning, i.e., neural network)
Predictors	Weekly averages from 1, 2, and 3 weeks ago of the squared daily inflow, weekly average inflow from 1 week ago	Current and prior water-year inflow	• Hydrometeorological variables: Air and water temperature, Hetch Hetchy (HH) inflow • System state variables: San Joaquin Pipeline flow (SJPL), HH storage, HH spill and release, Canyon Power Tunnel (CPT) • System memory: Flows and dry days in a previous water year, and in a recent water year, antecedent dry days, one-day-prior HH inflow
Variables and Locations	Turbidity at the O'Shaughnessy Dam	TOC at the Tesla Portal	Turbidity at the O'Shaughnessy Dam, turbidity, and TOC at the Tesla Portal
Relation	Linear	Linear	Nonlinear
Validation	Not included	Not included	Included
Time step	Weekly	Water year	Daily
Predictive uncertainty	Not available	Not available	Median prediction with 95% of confidence level

2 Data

The indicators of raw water quality (i.e., turbidity, TOC) are termed predictands. Turbidity at the O'Shaughnessy Dam has the most abundant data with 6496 days. Turbidity records at the Tesla Portal are 6767 days, while TOC at the Tesla Portal has 855 observed days. A sanity check is performed across the collected dataset to remove data records with measurement errors (e.g., outliers that are judged as 'errors' in agreement with SFPUC's personnel). Predictors are divided into three groups: the hydroclimate, the system state, and the system memory described in Table 3. Hydroclimatic predictors are measurements relating to hydrology and climate, such as the Hetch Hetchy inflow, the air temperature over the Hetch Hetchy Reservoir, the water temperature at the Tesla Portal. System state predictors (i.e., operational drivers) in the Hetch Hetchy water supply system are composed of the SJPL flow, the Hetch Hetchy storage, the Hetch Hetchy spill and release (i.e., Hetch Hetchy Reservoir spill and release into the Tuolumne River for undesired excess water and for maintaining instream flow requirement), and the Canyon Power Tunnel (CPT) flow (see Figure 2). Since the Hetch Hetchy Reservoir is a big lake with a storage capacity reaching 360,360 acre-feet (444.5 million cubic meters), according to Null and Lund (2006), we need to consider the reservoir's memory in operational and hydroclimatic variables. In particular, hydroclimatic and system state variables in previous time steps become system memory variables.

We classify system memory inputs into long-term memory, middle-term memory, and short-term memory. Long-term memory refers to measurements of the previous water year, including cumulative dry days and flows. The middle-term memory implies dry days and flows within the current water year, computed as accumulated values from the first date to the predicted date in the current water year. Short-term memory considers the one-day-prior Hetch Hetchy flow and Hetch Hetchy antecedent dry days (i.e., no-precipitation days accumulated from the last day having precipitation until the predicted date). SFPUC measures all predictors.

Table 2 Training and validation periods of turbidity at the O'Shaughnessy Dam, turbidity and TOC models at the Tesla Portal

		Training dataset	Validation dataset	Full dataset
Turbidity at O'Shaughnessy Dam	Mean (NTU)	0.36	0.25	0.34
	Standard deviation (NTU)	0.43	0.14	0.39
	Number of days	5414	1353	6767
	Duration	01/01/1996-12/24/2010	12/25/2010-09/30/2014	01/01/1996-09/30/2014
	Missing days	58	23	81
Turbidity at the Tesla Portal	Mean (NTU)	0.48	0.35	0.46
	Standard deviation (NTU)	0.23	0.13	0.22
	Number of days	5197	1299	6496
	Duration	01/01/1996-12/20/2010	12/21/2010-09/30/2014	01/01/1996-09/30/2014
	Missing days	271	81	352
TOC at the Tesla Portal	Mean (mg/L)	1.36	1.34	1.35
	Standard deviation (mg/L)	0.24	0.20	0.23
	Number of days	684	171	855
	Duration	01/03/1999-05/25/2011	05/29/2011-09/28/2014	01/03/1999-09/28/2014
	Missing days	3841	1047	4892

Table 3 List of predictors

	No.	Name (Unit)	Short name	Dimension	TOC at the Tesla Portal	Turbidity at the Tesla Portal	Turbidity at the O'Shaughnessy Dam
Hydroclimatic predictors	1	Hetch Hetchy inflow (Cubic feet per day, CFD)	Q_HH_cfd	Volume/Time	✓	✓	✓
	2	Hetch Hetchy air temperature (Degree Celsius, °C)	Ta_HH_C	Temperature	✓	✓	✓
	3	Tesla water temperature (Degree Celsius, °C)	Tw_Tesla_C	Temperature	✓	✓	
System state predictors	4	San Joaquin Pipeline flow (Millions of gallons per day, MGD)	SJPL_mgd	Volume/Time	✓	✓	
	5	Hetch Hetchy storage (Acre-foot, AF)	V_HH_af	Volume	✓	✓	✓
	6	Hetch Hetchy spill and release (Cubic foot per second, CFS)	SR_HH_cfs	Volume/Time	✓	✓	✓
	7	Canyon Power Tunnel flow (Cubic foot per second, CFS)	CPT_cfs	Volume/Time	✓	✓	
Long-term memory	8	Cumulative dry days in the previous water year (Days)	ADD_HH_sum1	Day	✓	✓	✓
	9	Cumulative flows in the previous water year (Cubic foot, CF)	Q_HH_cfd_sum1	Volume	✓	✓	✓
Middle-term memory	10	Cumulative dry days until the predicted day in the current water year (Days)	ADD_HH_sum	Day	✓	✓	✓
	11	Cumulative flows until the predicted day in the current water year (Cubic foot, CF)	Q_HH_cfd_sum	Volume	✓	✓	✓
Short-term memory	12	Hetch Hetchy antecedent dry days (days)	ADD_HH	Day	✓	✓	✓
	13	One-day-prior Hetch Hetchy inflow (Cubic feet per day, CFD)	Q_HH_cfd1	Volume			✓

3 Methodology

3.1 Composite Quantile Regression Neural Network

3.1.1 Mechanism

The availability of remote sensing products and long records of ground measurements has motivated machine learning approaches in geosciences and engineering. Water quality is a young science with a lot of undiscovered knowledge (Costa et al. 2021). Data from measurements is sufficient for artificial intelligence models to investigate current phenomena (Najah et al. 2013). Among these techniques, ANN has been widely applied in water quality modeling (Tiyasha, Tung, and Yaseen 2020) due to its ability to represent complex relationships among many processes (Sinshaw et al. 2019).

Figure 6 illustrates the structure of the ANN model. The network contains three layers: the input layer, the hidden layer, and the output layer. Figure 6 shows the network with only one layer as an example, although more layers can be considered in ANN. The value H_j at the hidden layer neuron, j for the time step t is obtained by applying the activation function g_1 to the weighted sum of the input x_i:

$$H_{j,t} = \sum_{i=1}^{I} g_1(w_{i,j}x_{i,t} + b_j), \quad 1$$

with I being the number of inputs (i.e., predictors), $w_{i,j}$ the weight associated with the input x_i and the hidden layer neuron j, b_j the hidden-layer bias.

Similarly, the value of the output neuron (i.e., the predictand) is obtained by applying the activation function g_2 to the weighted sum of the output from each hidden layer neuron H_j:

$$\hat{y}_t = \sum_{j=1}^{J} g_2(v_j H_{j,t} + c), \quad 2$$

where J is the number of neurons in the hidden layer and v_j the weight associated with the value H_j of the hidden layer neuron j, $\hat{y}$ the predicted value of the considered predictand at the time step t, c is the output-layer bias.

After combining equation 1 and 2, the general function is summarized in equation 3:

$$\hat{y}_t = \sum_{j=1}^{J} g_2(v_j \sum_{i=1}^{I} g_1(w_{i,j}x_{i,t} + b_j) + c), \quad 3$$

Predicted values ($\hat{y}_t$) of ANN should be close to observed values to produce a good predictive performance. ANN needs the training (i.e., the calibration) to achieve that. Training or calibration is the ANN's learning process by modifying parameters, including the weights $w_{i,j}$, and v_j to minimize errors (i.e., cost or loss or objective functions) between predicted and observed values (Goh 1995). Throughout the training, ANN's parameters (i.e., weights $w_{i,j}$ and v_j and biases b_j and c) are updated through a backpropagation procedure that uses a stochastic gradient-based nonlinear optimization (Cannon 2011) to minimize the error function. Once the weights are updated, a new prediction can be made, and further evaluation of the error function is done. This procedure is repeated until the convergence of the error function.

ANN inputs are first standardized to zero mean and unit standard deviation (Cannon 2011), facilitating the convergence of the backpropagation procedure.

The logistic function, a typical activation function in neural network's applications (Chau 2006), is used for $g_1$and g_2:

$$g(n) = \frac{1}{1 + e^{-n}}. \quad (4)$$

Other activation functions are also commonly used, such as the Rectified Linear Unit (ReLU), and the Hyperbolic Tangent (Tanh) functions.

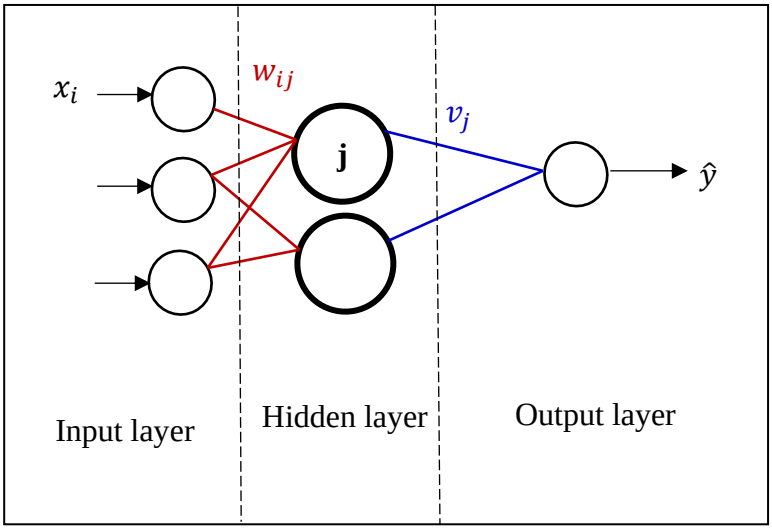

Figure 6 Artificial Neural Network structure. x_i is the input; w_{ij} is the weight between input i and neuron j; v_j is the weight between neuron j and predicted output $\hat{y}$

The ANN model discussed above is meant to be calibrated against an error measure, usually mean squared error, between predicted and observed values; hence, it is appropriate to normally distributed data (Maier and Dandy 2000) and aims to predict the mean of the predictands instead of their percentiles. In other words, ANN is not designed to estimate skewed data and tails of the distribution (Taylor 2000), implying its weakness for simulating high and low values in the dataset. Hence, Quantile Regression Neural Network (QRNN) is created to overcome this challenge. Q. Xu et al. (2016) found that QRNN is more robust than local linear regression and spline regression

in fitting outliers. QRNN combines the Quantile Regression method with the Artificial Neural Network (Taylor 2000). The main idea of QRNN is that the error function (denoted as *loss* or *loss function*) is built by quantile regression cost function, while predicted values are derived from a neural network (Taylor 2000) (see Equation 5). It helps model a chosen quantile of the predictand distribution rather than median or mean. The loss function to be minimized reads as (Taylor 2000):

$$E_{QRNN} = \sum_{t|y_t \geq \hat{y}_t} \tau|y_t - \hat{y}_t| + \sum_{t|y_t < \hat{y}_t} (1-\tau)|y_t - \hat{y}_t|, \quad 5$$

With E the loss function, τ the quantile of the predictand distribution to be predicted by the network, y_t the value of the predictand at the time step *t* and $\hat{y}_t$ the predicted quantile value of the predictand distribution. Note that $\hat{y}_t$ is obtained following equation 3. Note also in equation 5 that for values of τ larger than 0.5 (i.e., prediction focuses on values greater than the median of the distribution), values $\hat{y}_t$ that underestimate y_t contributed more to the loss function (i.e., weight equals $\tau > 0.5$) while values $\hat{y}_t$ that overestimate y_t contribute less (weight equals $1 - \tau < 0.5$). Conversely, if $\tau < 0.5$, a higher weight is given to predicted values $\hat{y}_t$ that underestimate y_t.

Nevertheless, QRNN still encounters crossing quantiles because quantile predictions are calibrated separately, leading to different parameter sets (Cannon 2018). To resolve this problem, Xu et al. (2017) and Cannon (2018) proposed Composite Quantile Regression Neural Network (CQRNN) that could simultaneously calibrate all quantile estimations. Consequently, weights and biases are unique for all quantile estimations (Xu et al. 2017, and Cannon 2018). The CQRNN algorithm has two benefits: 1) avoiding the crossing issue of quantile simulations (Cannon 2018), and 2) combining strengths from all quantile predictions to produce multiple quantile estimations close to true conditional quantile functions (Xu et al. 2017, and Cannon 2018). CQRNN has the same structure as QRNN. The only difference is that the loss function of CQRNN is the average from all quantile predictions (equation 6); whereas, the QRNN's error function merely considers the loss function for each quantile (equation 5). The separate calibration for each quantile in QRNN induces distinctive sets of weights and biases, which causes the overpassing of quantiles. According to Xu et al. (2017) and Cannon (2018), the loss function of CQRNN has the following form:

$$E_{CQRNN} = \frac{1}{K}\sum_{k=1}^{K} E_{QRNN}(\tau_k) = \frac{1}{K}\sum_{k=1}^{K}\left(\sum_{t|y_t \geq \hat{y}_t} \tau_k|y_t - \hat{y}_t| + \sum_{t|y_t < \hat{y}_t} (1-\tau_k)|y_t - \hat{y}_t|\right) \quad 6$$

With k = 1, 2... K are all desired quantiles, $E_{QRNN}(\tau_k)$ is the loss function of QRNN for a specific τ_k, described in equation 5. In this research, five quantiles (i.e., 2.5th, 25th, 50th, 75th, and 97.5th) are used.

3.1.2 Calibration

Figure 7 illustrates the calibration process, including training and validation for a one-layer neural network, considered for the raw water quality model of the Hetch Hetchy system (described

below). The training chunk (i.e., the calibration) includes the first 80% of the available time series of predictors/predictands. The validation chunk contains 20% of the remaining data (i.e., the time series's last 20% data points). In this study, the validation size is 25% of training, which complies with Najah et al. (2013) that this ratio should be 10-40%. A good training set should include all extreme events (Najah et al. 2013). Thus, the training dataset in this study consists of the largest predictand values, such as the highest turbidity value ever recorded at the Tesla Portal in 1997, which ensures the network to be trained to reproduce such critical events. Details of training and validation datasets for turbidity and TOC are described in Table 2.

The ensemble technique can improve a single neural network (Khalil, Ouarda, and St-Hilaire 2011). This method combines results from neural networks to produce the ensemble output (Khalil, Ouarda, and St-Hilaire 2011). Several ways are used to generate ensemble neural networks, such as using 1) different initial random weights, 2) diverse network topology, 3) varied training algorithms, and 4) divergent training sets (Khalil, Ouarda, and St-Hilaire 2011). Among them, manipulating training datasets are popular (Khalil, Ouarda, and St-Hilaire 2011). *Bagging* is the most frequently used (Khalil, Ouarda, and St-Hilaire 2011). Therefore, the training set is split into five sub-datasets using the *bagging* method (Cunningham, Carney, and Jacob 2000). This method consists of creating N training subsets by randomly sampling observed time series of predictors and predictands with replacement from the original training set (Figure 7). The objective is to minimize the dependence of model parameters on the training set by creating new training sets that are similar but still different from the original so that the risk of reaching a local optimum in the parameter space is reduced (Dreiseitl and Ohno-Machado 2002). As shown in Figure 7, five *bags* are used in this study (i.e., N=5) as a good compromise between the computation's efficiency and reduced risk of reaching a local optimum in the parameter space. Some papers (Shu and Ouarda 2007; and Khalil, Ouarda, and St-Hilaire 2011) stated that the ensemble should be averaged from at least five neural networks (i.e., five bags) to minimize the estimated error. The final prediction of CQRNN is generated from the ensemble across the five CQRNN models (i.e., noted as *Ensemble CQRNN* in Figure 7). Furthermore, each CQRNN is optimized from five repeated runs with different starting weights and biases to avoid the local minima of the loss function. This study uses two methods to create the neural network ensemble for improving the model's robustness: different randomly initial conditions (i.e., weights and biases) and training sets (Khalil, Ouarda, and St-Hilaire 2011).

The number of neurons in the hidden layer (only one hidden layer is used in this study) is chosen so that the averaged loss function across five CQRNNs is minimum for the validation set. As explained in the mechanism above, each CQRNN model (optimized from five different random sets of weights and biases) has the error function, averaged from five quantiles: 2.5^{th}, 25^{th}, 50^{th}, 75^{th}, and 97.5^{th}. For calibration, the optimal number of neurons is selected by the least loss function averaged from five bagged samples and five quantiles (i.e., 2.5^{th}, 25^{th}, 50^{th}, 75^{th}, and 97.5^{th}) in the validation chunk. The selection of hyperparameters (i.e., the number of neurons in this study) from validation prevents CQRNN from overfitting. The more neurons are, the lower losses are in the calibration (see Figure 8, Figure 9, and Figure 10). However, too many neurons provoke the

overfitting, while too few neurons evoke the underfitting (Khalil, Ouarda, and St-Hilaire 2011). Therefore, hyperparameters of CQRNN should be chosen from the validation dataset, which is not used by the training. The CQRNN model is built in the R programming language and package *mcqrnn* (Cannon 2019).

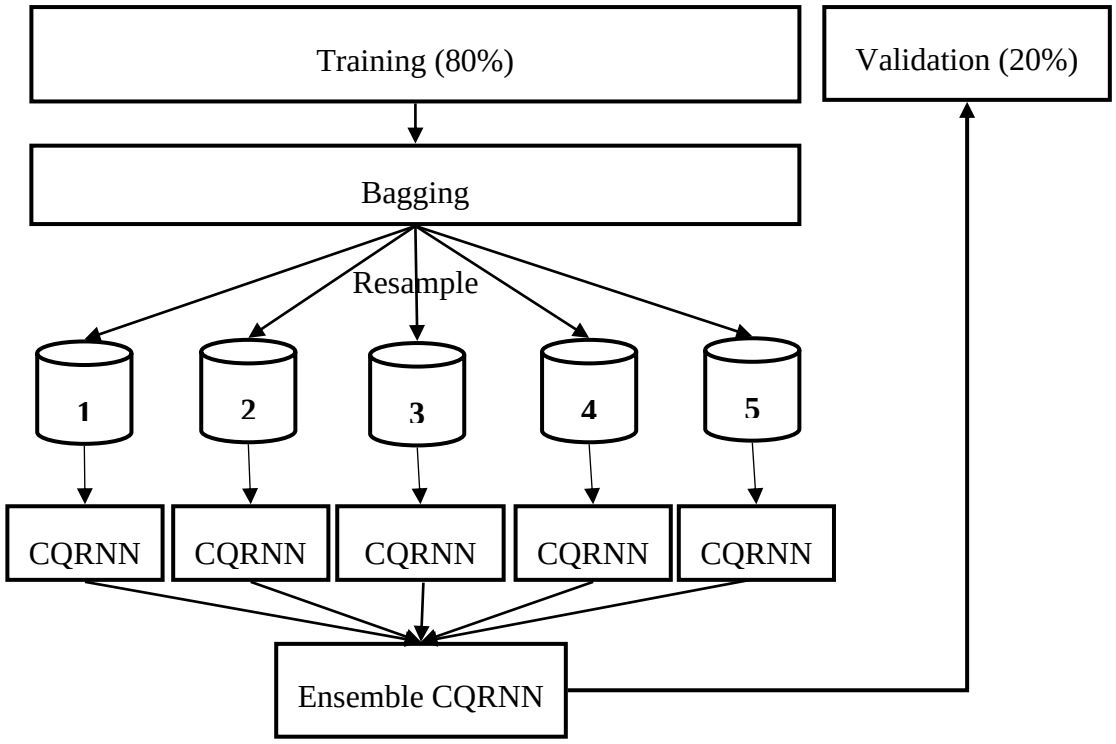

Figure 7 Bagging setting in this study. CQRNN is the Composite Quantile Regression Neural Network (optimized from five random sets of weights and biases)

3.2 Input determination for the CQRNN approach

For the selection of predictors, we combine three methods which are i) priori knowledge, ii) cross-correlation, iii) heuristic (trial and error) approaches (Bowden, Dandy, and Maier 2005). The prior knowledge is using the expertise to define candidate predictors for water quality prediction. Cross-correlation means that we remove variables that are highly correlated to each other. The heuristic approach implies that we build several water quality models from several combinations of predictors and chose the best model, which has the best predictive performances. As a result, we find that turbidity and TOC models at the Tesla Portal need all twelve predictors listed in Table 3. We use the one-day-prior HH inflow for turbidity at the O'Shaughnessy Dam because the highest HH inflow caused by a rain-on-snow event (01/02/1997) occurred one day before high turbidity spikes (01/03/1997 - 01/15/1997). This phenomenon is caused by several reasons such as the difference in the sampling of observations, dam operations, etc. For turbidity and TOC at the Tesla Portal, we do not use the one-day-prior HH inflow since this location is indirectly affected by this predictor. Moreover, the turbidity model at the O'Shaughnessy Dam excludes the Canyon Power Tunnel flow and the San Joaquin Pipeline flow because the Canyon Power Tunnel and the San Joaquin Pipeline are located downstream of the dam. Table 3 shows the predictors used in each model.

3.3 Predictors' relative importance

Artificial neural networks are often mentioned as a 'black box' linking a set of predictors to predicting a chosen variable. However, several methods exist to open the 'black box', which allows to understand better the contribution of each predictor to the output of the network (e.g., Pentoś 2016). In this study, we use the connection weight method as described in (Gevrey, Dimopoulos, and Lek (2003) to estimate the contribution of each predictor to the network prediction. A summary of this method is provided below.

As shown in Figure 6, each predictor *i* value to be input to a given neuron *j* of the hidden layer is assigned a weight w_{ij} (equation 1). In addition, the output from each neuron *j* of the hidden layer is assigned a weight v_j. According to the connection weight method (Gevrey, Dimopoulos, and Lek 2003), these weights can represent the contribution from each predictor to the final network prediction. They are somehow similar to coefficients (β) in the equation 8.

The contribution from each predictor is further discussed using the Relative Importance (RI) metric first proposed by Garson (1991) described in Gevrey, Dimopoulos, and Lek (2003); Kemp, Zaradic, and Hansen (2007); (de Oña and Garrido 2014):

$$RI_i = \frac{\sum_{j=1}^{J} |w_{ij}| \times |v_j|}{\sum_{i=1}^{I} \sum_{j=1}^{J} |w_{ij}| \times |v_j|} \times 100 \qquad 7$$

Where RI_i is the relative importance of the predictor i (%), w_{ij} is the weight assigned to the predictor i and neuron j, v_j is the weight assigned to the output of the neuron j, J is the number of hidden neurons, and I is the number of predictors.

3.4 Other Data-driven Methods

To evaluate the performance of the developed approach, CQRNN is compared with other data-driven methods commonly used to predict water quality such as QR (e.g., Francke, López-Tarazón, and Schröder 2008), LR (e.g., Francke, López-Tarazón, and Schröder 2008), MARS (Heddam and Kisi 2018), and KNN (Lee and Scholz 2006). We use identical training and validation datasets for all methods to make a fair comparison. In other words, only one set of predictors is used in all five algorithms. Below, a short description of each model is provided. More details for each technique are given in the cited literature.

3.4.1 Linear Regression

LR is a parametric approach to fit a linear relationship between predictors and a predictand. The general form of a multivariate LR is as follows:

$$\hat{y}_t = \beta_0 + \beta_1 \times x_{1,t} + \cdots + \beta_n \times x_{I,t} + \varepsilon_t \quad 8$$

Where $\hat{y}_t$ the predicted value of the predictand at the time step t, x_i, t the value of the predictor i at the time step t and I the number of predictors (i.e., inputs), β_i the regression coefficient for the predictor i, ε_t an error term. The regression coefficients β_i are optimized to minimize the mean square error (i.e.,$\sum_t {\varepsilon_t}^2$). The LR model is built based on the function *lm* from the *stats* R package (Wilkinson and Rogers 1973).

3.4.2 Quantile Regression

QR (Koenker and Bassett 1978) consists of fitting a LR model using the loss function described in equation 5 by substituting $\hat{y}_t$ by equation 8. The QR model is built by using the function *rq* from the *quantreg* R package (R. Koenker et al. 2019).

3.4.3 Multivariate Adaptive Regression Spline

MARS is a nonparametric method (Friedman 1991). It is useful when predictors and predictand have a nonlinear relationship. If not, MARS is similar to LR. When nonlinearities exist, such as a change in slope at a specific location in the predictor space (commonly called *knot*), the MARS algorithm identifies the knot. It then fits a LR using the subsets of predictors on each side of the knot. The predicted value $\hat{y}_t$ is then given as:

$$\hat{y}_t = \sum_{k=1}^{K} l_k(\hat{y}_{k,t}) + \varepsilon_t, \quad (9)$$

Where $\hat{y}_{k,t}$ is the predicted value obtained from the regression between the knot *k* and *k+1*, *K* is the number of knots and l_k is the Hinge function that returns identity if below the knot *k* within the predictor space, and 0 otherwise. Note that the MARS algorithm defines the optimal number of knots to minimize the squared error. In equation 9, $\hat{y}_{k,t}$ is obtained using equation 8 across the subset of predictors located between the knot *k* and *k+1*. The MARS model is built using the *mars* function from the *mda* R package (Leisch and Ripley 2017).

3.4.4 K-Nearest Neighbors

KNN is a nonparametric resampling algorithm (Lall and Sharma 1996). KNN prediction is based on the weighted average of predictand values for *K* closest *neighbors*. In this context, the *K* is a hyper-parameter. The *neighbors* are predictand values observed during dates for which predictors are similar to the current time simulation step. The weight given to each selected predictand is the averaged inverse of the Euclidean distance between predictors. A larger weight is given to predict and for which predictors are more similar to the current simulated time step. The KNN sampling approach is built based on using the function *train* from the *caret* R package (Wing et al. 2019).

3.5 Predictive Performance Measures

3.5.1 Loss Function of Composite Quantile Regression Neural Network

Since CQRNN calibrates model parameters through minimizing a *loss function*, we first evaluate CQRNN by this loss. The loss should be small enough to confirm a good model. The loss function is expressed by equation 6.

3.5.2 Absolute Error (AE)

Absolute Error (AE) can detect the averaged bias between the prediction and the observation, revealing the performance of models. AE should be as small as possible. AE is calculated as follows.

$$AE = \left| \frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)}{T} \right| \quad (10)$$

Where T total number of time steps, y_t the observed value at time step t, $\hat{y}_t$ the predicted value at time step t. In our paper, the time step is daily; the value is water quality (specifically, turbidity and TOC for the case study).

4 Results and Discussion

To reveal the performance of the water quality model and predictors' effect, the result section is composed of three main sub-sections. Section 4.1 presents results from selecting the number of neurons for each CQRNN model (denoted as a hyper-parameter). The number of hidden layers (see Figure 6) is also referred to as a hyper-parameter, but we only use one layer in this study. Therefore, the selection is only conducted for the number of neurons. Section 4.2 shows the comparison among CQRNN and other conventional methods (QR, LR, MARS, and KNN) in simulating turbidity and TOC at the Tesla Portal, turbidity at the O'Shaughnessy Dam. Section 4.3 aims to define significant explanatory variables on water quality.

4.1 Calibration of hyper-parameters in CQRNN

As the number of neurons in the hidden layer increases (i.e., the model complexity increases), the value for the loss function obtained for the training set commonly decreases. For example illustrated in Figure 8, Figure 9, and Figure 10, where the values of the loss function obtained for the training in turbidity and TOC models decrease as the number of neurons increases. Parsimonious models (i.e., a model with a small number of parameters) are usually preferred to prevent overfitting the model parameters. In this context, the overfitting of the model parameters often increases the loss function for the validation set with the number of parameters. For this reason, the number of neurons for each location is selected as the one that minimizes the value of the loss function obtained for the validation set. The optimal numbers of neurons for each model are listed in Table 4.

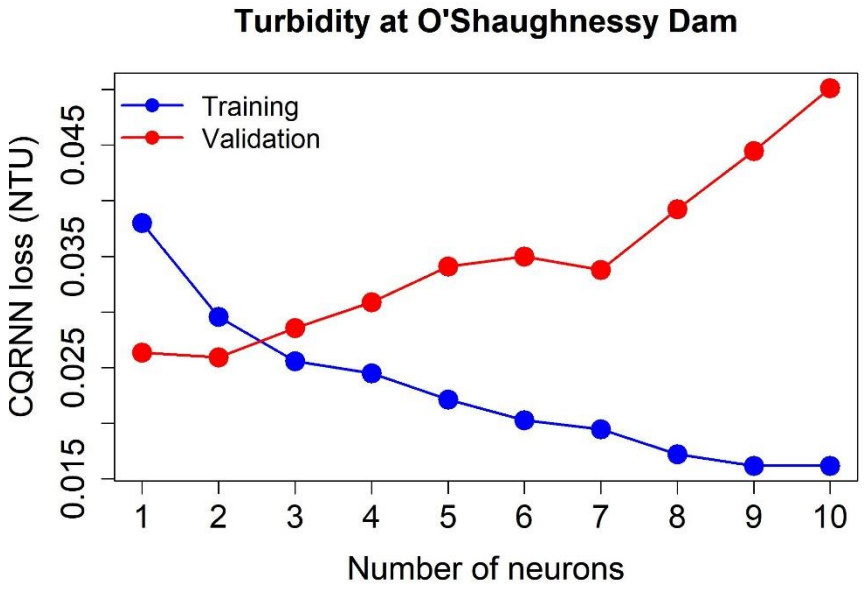

Figure 8 Evolution of the CQRNN losses (averaged quantile regression errors over 2.5th, 25th, 50th, 75th, 97.5th percentiles, described in equation 6) with the number of neurons in one hidden layer for turbidity at the O'Shaughnessy Dam

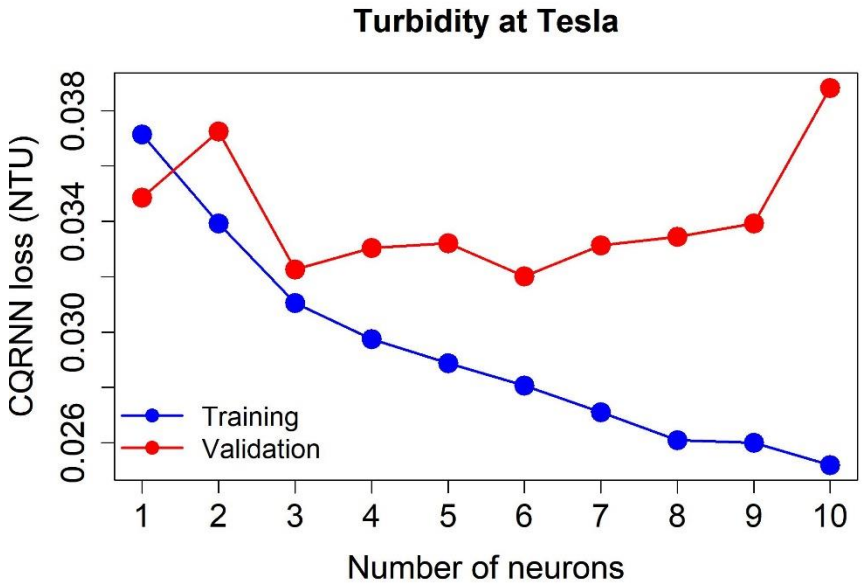

Figure 9 Evolution of the CQRNN losses (averaged quantile regression errors over 2.5th, 25th, 50th, 75th, 97.5th percentiles, described in equation 6) with the number of neurons in one hidden layer for turbidity at the Tesla Portal

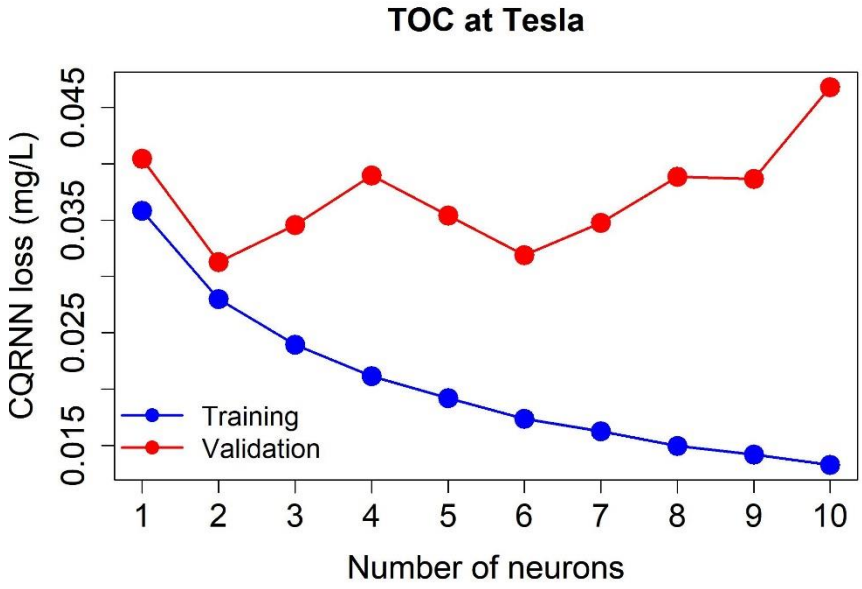

Figure 10 Evolution of the CQRNN losses (averaged quantile regression errors over 2.5th, 25th, 50th, 75th, 97.5th percentiles, described in equation 6) with the number of neurons in one hidden layer for TOC at the Tesla Portal

Table 4 Optimal number of neurons for water quality models in the Hetch Hetchy Regional Water System

	Turbidity	Total organic carbon (TOC)
The O'Shaughnessy Dam	2	
The Tesla Portal	6	2

4.2 Performances of CQRNN in water quality prediction in the Hetch Hetchy Regional Water System

To evaluate the general predictive performance of models, the loss (see equation 6), and the AE (see equation 10) are investigated. The loss is the optimizer of CQRNN and QR, while AE is similar to the mean squared error, which is the objective function in the LR's optimization. AE is used because it has the same unit as the loss function. Losses and AEs from CQRNN, QR, LR, MARS, and KNN for all sites and predictands are listed in

Table 5. These metrics are visually illustrated in Figure 11, Figure 12, and Figure 13, respectively for turbidity at the O'Shaughnessy Dam, turbidity, and TOC at the Tesla Portal. In each model, the loss is computed as the average over five quantiles (i.e., 2.5^{th}, 25^{th}, 50^{th}, 75^{th}, 97.5^{th} percentiles) (see equation 6). For CQRNN and QR, five quantile estimations are distinct. However, LR, MARS, and KNN are identical because these methods generate only one mean prediction. Therefore, predicted and observed values are the same for five quantiles in equation 6 when computing losses for LR, MARS, and KNN. The only change is the quantile values τ in equation 6. AEs are calculated based on median estimations in CQRNN and QR (i.e., median is a proxy for the mean in these methods). AEs are also obtained for LR, MARS, and KNN, which produce mean predictions.

In

Table 5, red numbers indicate the best models. All losses and AEs from the CQRNN model are minimums, except for training AEs. This fact reveals the superior performance of the CQRNN model over QR, LR, MARS, and KNN. The KNN approach seems to be the second-best option for water quality modeling based on the lowest training AE. However, KNN is not good because of poor performances in validation AEs. The comparison among five methods (i.e., CQRNN, QR, LR, MARS, and KNN) is impartially conducted because five models use the same bagging technique. CQRNN performs well in both loss and AE, although CQRNN does not optimize AE; this suggests that CQRNN's median predictions are close to the mean. Other methods do not perform well for error functions that they do not optimize for. For example, LR, MARS, KNN lack quantile estimations so that their predictive performances are understandably poor for losses, which are averaged over five percentile predictions. Unlike CQRNN, QR performs poorly in terms of the assessment based on AEs since they are not the cost functions in this technique.

Table 5 Predictive performance measures as losses (quantile regression errors from equation 6) and absolute errors (AEs) of CQRNN (Composite Quantile Regression Neural Network), QR (Quantile Regression), LR (Linear Regression), MARS (Multivariate Adaptive Regression Splines), KNN (K-Nearest Neighbors) for turbidity at the O'Shaughnessy Dam (Turb_OSH), turbidity and TOC at the Tesla Portal (Turb_Tesla and TOC_Tesla). Training and validation sets are described in Table 2. The training dataset accounts for 80% of the full data, while the remaining 20% data is used for validation (see Table 2). The training data has a larger variability and more extremes than the validation one (see Table 2). Red numbers are the least losses and AEs, implying the best model

			CQRNN	QR	LR	MARS	KNN
Loss	Turb_OSH (NTU)	Training	0.030	0.038	0.069	0.061	**0.008**
		Validation	**0.026**	**0.026**	0.061	0.059	0.064
	Turb_Tesla (NTU)	Training	**0.028**	0.037	0.063	0.056	0.032
		Validation	**0.032**	0.033	0.067	14.71	0.073
	TOC_Tesla	Training	**0.028**	0.034	0.060	0.045	0029
	(mg/L)	Validation	**0.031**	0.041	0.069	0.079	0.077
Absolute error	Turb_OSH (NTU)	Training	0.227	0.391	0.368	0.346	**0.099**
		Validation	0.136	**0.125**	0.176	0.160	0.172
	Turb_Tesla (NTU)	Training	0.168	0.202	0.197	0.170	**0.130**
		Validation	**0.137**	0.142	0.170	45.10	0.188
	TOC_Tesla	Training	0.127	0.158	0.156	0.121	**0.086**
	(mg/L)	Validation	**0.131**	0.174	0.174	0.201	0.189

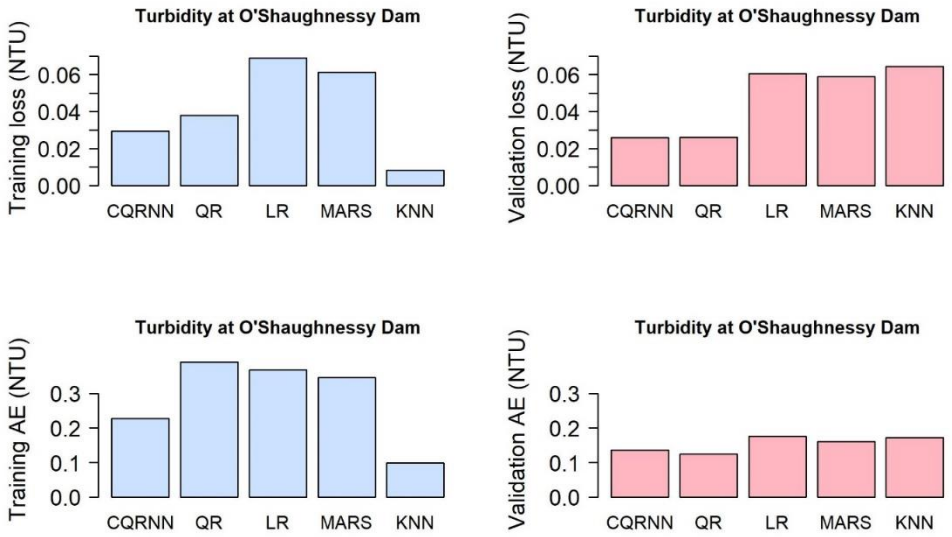

Figure 11 Compare the performance of CQRNN (Composite Quantile Regression Neural Network) with QR (Quantile Regression), LR (Linear Regression), MARS (Multivariate Adaptive Regression Splines), KNN (K-Nearest Neighbors) for turbidity at the O'Shaughnessy Dam by losses (quantile regression errors from equation 6) and absolute errors (AEs). Top left, top right, bottom left, and bottom right figures respectively illustrate training loss, validation loss, training AE, and validation AE

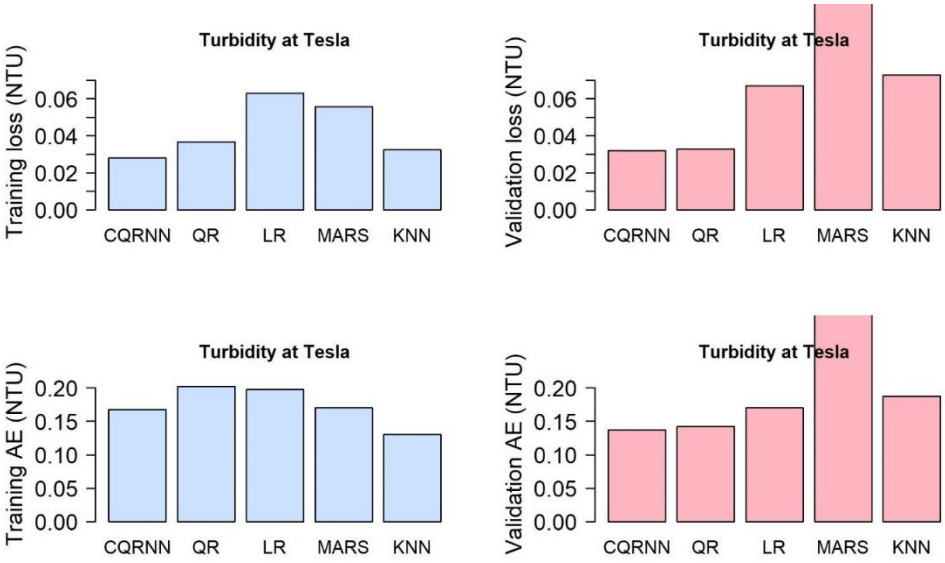

Figure 12 Compare the performance of CQRNN (Composite Quantile Regression Neural Network) with QR (Quantile Regression), LR (Linear Regression), MARS (Multivariate Adaptive Regression Splines), KNN (K-Nearest Neighbors) for turbidity at the Tesla Portal by loss (quantile regression error from equation 6) and absolute error (AE). Top left, top right, bottom left, and bottom right figures respectively illustrate training loss, validation loss, training AE, and validation AE

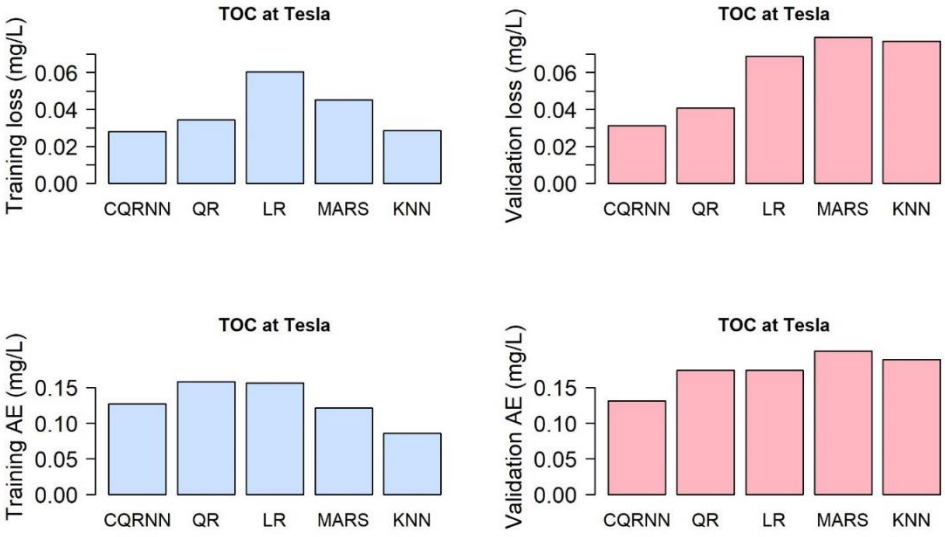

Figure 13 Compare the performance of CQRNN (Composite Quantile Regression Neural Network) with QR (Quantile Regression), LR (Linear Regression), MARS (Multivariate Adaptive Regression Splines), KNN (K-Nearest Neighbors) for TOC at the Tesla Portal by loss (quantile regression error from equation 6) and absolute error (AE). Top left, top right, bottom left, and bottom right figures respectively illustrate training loss, validation loss, training AE, and validation AE

4.2.1 Turbidity at the O'Shaughnessy Dam

For water managers, the exceedance of water quality levels over the regulatory limit is crucial. Therefore, a water quality model, which can predict extreme events, is as helpful as performing well on the data mean. A rain-on-snow event caused high turbidity values at the O'Shaughnessy

Dam and the Tesla Portal in January 1997. Thus, the performances of turbidity models at these locations should be assessed by the ability of re-producing turbidity peaks in January 1997. Figure 14 illustrates the rain-on-snow event in 1997, generating high inflow (middle right) and storage (bottom right) at the Hetch Hetchy Reservoir. This event triggered high turbidity values (top right of Figure 14) and the SJPL shutdown. CQRNN (top left of Figure 14) and KNN (bottom left of Figure 14) can capture high turbidity levels between January 3-15, 1997. However, as the interpretation for

Table 5, KNN has a poor performance on the validation dataset. Therefore, KNN is not selected due to its overfitting. LR, MARS (bottom left of Figure 14), and QR (middle left of Figure 14) fail to simulate turbidity peaks between January 3-15, 1997. In brief, the CQRNN's performance is the most satisfactory due to its capability to simulate both variability and turbidity peaks at the O'Shaughnessy Dam.

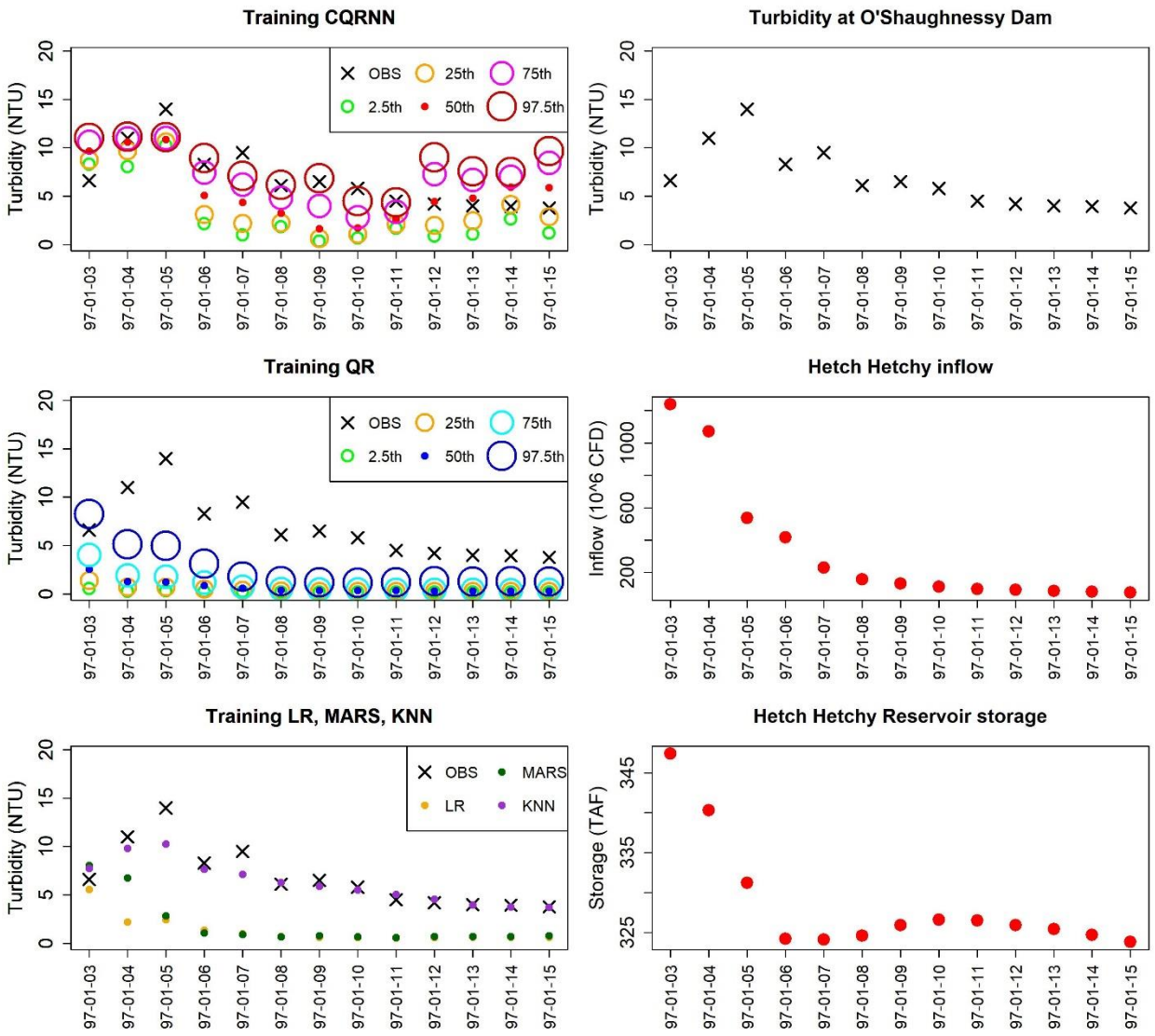

Figure 14 Compare the performance of CQRNN (Composite Quantile Regression Neural Network) with QR (Quantile Regression), LR (Linear Regression), MARS (Multivariate Adaptive Regression Splines), KNN (K-Nearest Neighbors) in modeling turbidity peaks at the O'Shaughnessy Dam in January, 1997. The comparison is illustrated by the simulation of a rain-on-snow event causing 13 high turbidity days (01/03/1997 - 01/15/1997) at the O'Shaughnessy Dam in CQRNN (Composite Quantile Regression Neural Network), QR (Quantile Regression), LR (Linear Regression), MARS (Multivariate Adaptive Regression Spline), KNN (K-Nearest Neighbors). Left figures from top to bottom respectively illustrate CQRNN, QR, and three methods from LR, MARS, KNN. In top left and middle left figures, "2.5th", "25th", "50th", 75th", "97.5th" are respectively 2.5th, 25th, 50th, 75th, 97.5th percentile predictions. Right figures from top to bottom respectively show the observed turbidity levels (OBS), the Hetch Hetchy inflow, and the Hetch Hetchy Reservoir storage (in thousands of acre-feet, TAF) during these 13 days

The performance of CQRNN in modeling turbidity at the O'Shaughnessy Dam is revealed in Figure 15 and Figure 16, respectively, by plotting time series and cumulative density functions (CDFs). Being similar to Figure 14, it is found that CQRNN can simulate high turbidity events in 1997 based on illustrations from Figure 15 and Figure 16. Furthermore, thanks to the quantile estimation, observed turbidity values at the O'Shaughnessy Dam are fallen within the 95% predictive range. Therefore, CQRNN and QR can reach more high and low turbidity levels than

LR, MARS, and KNN (see Figure 15 and Figure 16). Particularly, when looking at cumulative density curves in Figure 16, observations are captured by CQRNN and QR, but they are uncaptured by LR, MARS, and KNN in the validation period.

Moreover, it is observed that the 95% prediction interval from QR, generated by both 2.5^{th} and 97.5^{th} percentile estimations, is wider than the one of CQRNN based on time series in Figure 15 and CDFs from Figure 16. Thus, CQRNN has lower underestimated and over estimated errors from 2.5^{th} and 97.5^{th} percentile models than QR. In general, CQRNN better simulates the variability and peaks of turbidity data at the O'Shaughnessy Dam than other models in both training and validation.

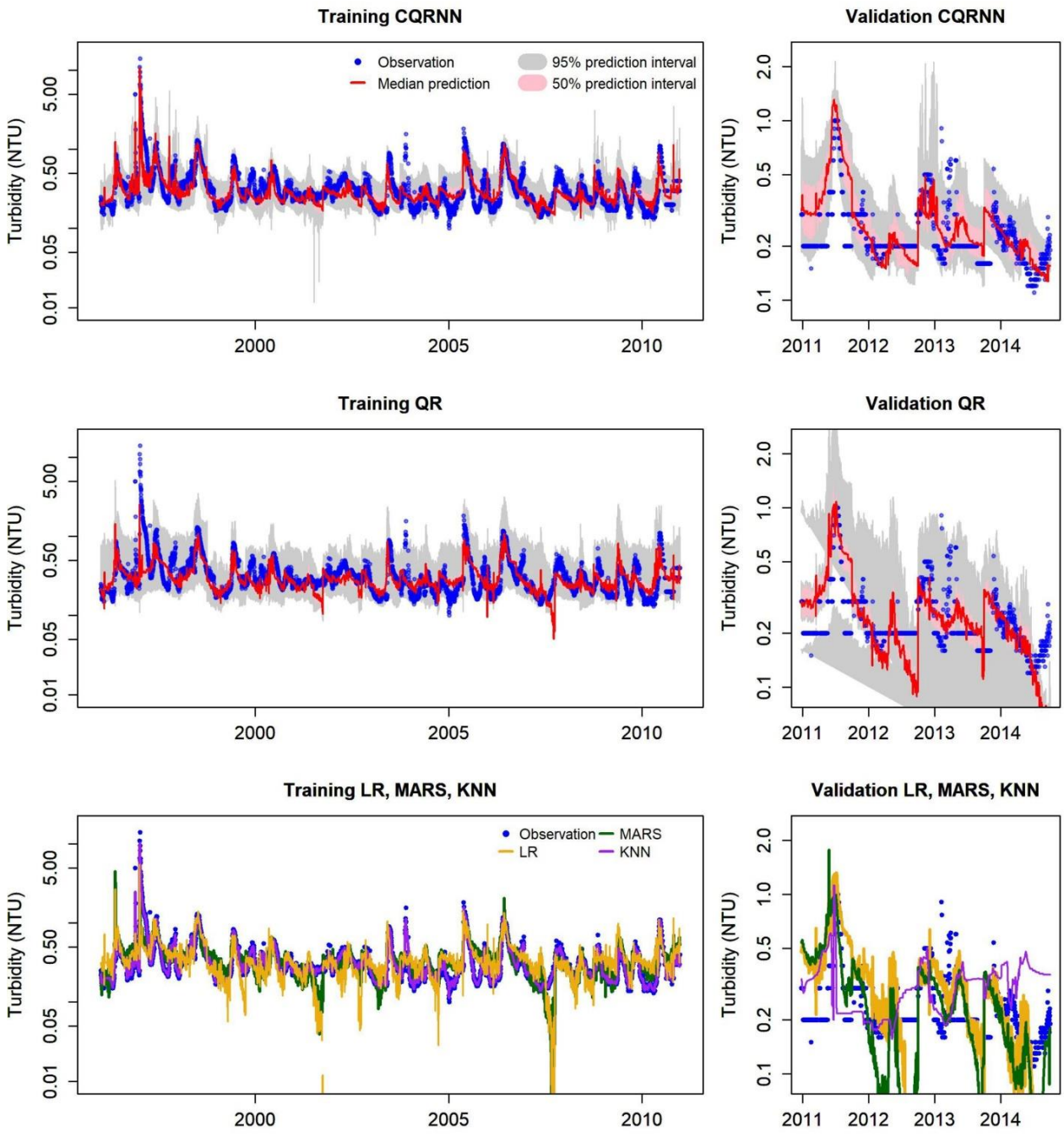

Figure 15 Predicted and observed time series of turbidity at the O'Shaughnessy Dam in training and validation periods for five methods: CQRNN (Composite Quantile Regression Neural Network), QR (Quantile Regression), LR (Linear Regression), MARS (Multivariate Adaptive Regression Splines), KNN (K-Nearest Neighbors). Top left, top right, middle left, middle right, bottom left and bottom right respectively illustrate training period of CQRNN, validation period of CQRNN, training period of QR, validation period of QR, training periods of LR, MARS, KNN, and validation periods of LR, MARS, KNN, The 50% predictive interval is the area between the 25th and 75th percentile models. The 95% predictive interval is the area between the 2.5th and 97.5th percentile models

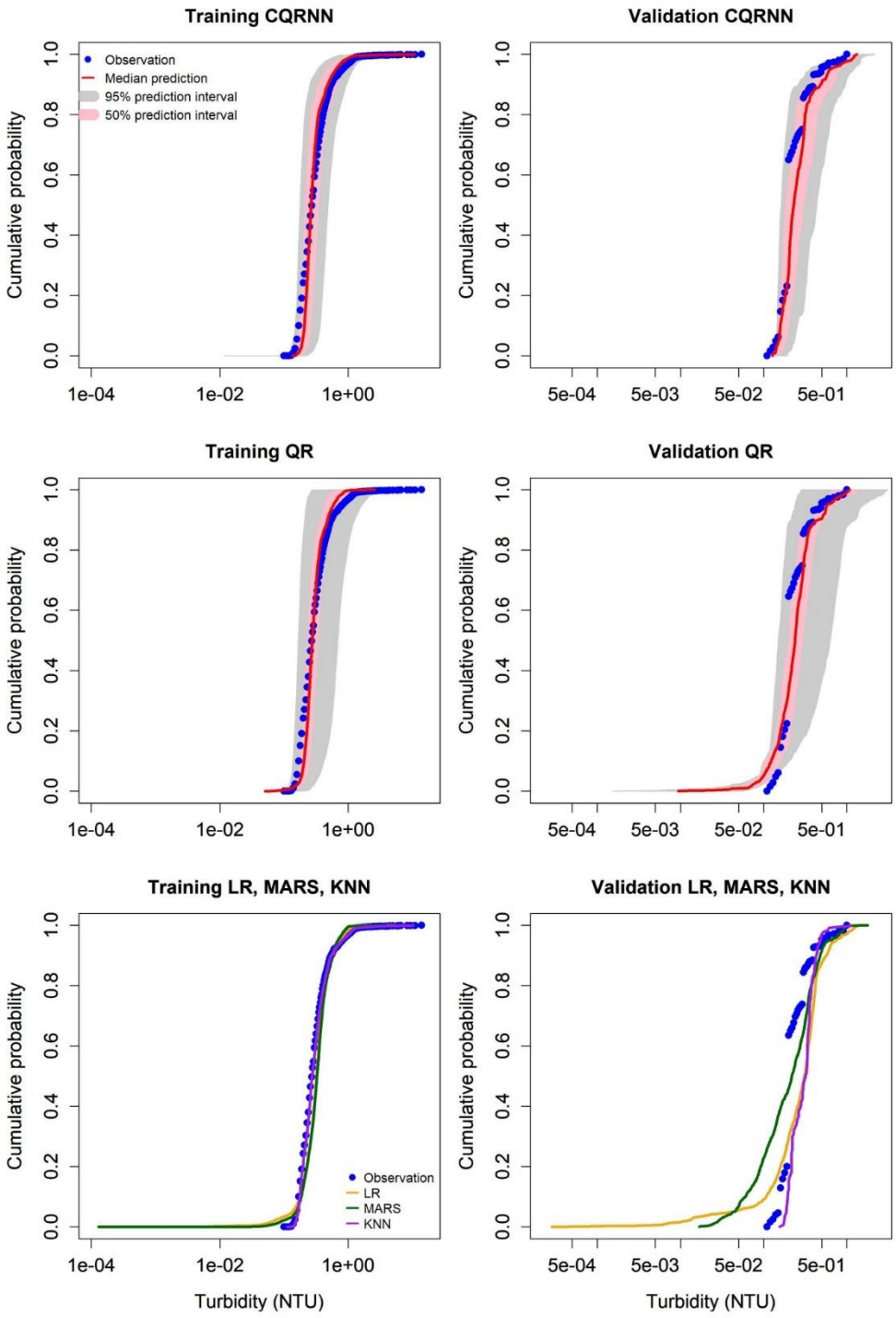

Figure 16 Predicted and observed cumulative density functions of turbidity at the O'Shaughnessy Dam in training and validation periods for five methods: CQRNN (Composite Quantile Regression Neural Network), QR (Quantile Regression), LR (Linear Regression), MARS (Multivariate Adaptive Regression Splines), KNN (K-Nearest Neighbors). Top left, top right, middle left, middle right, bottom left and bottom right respectively illustrate training period of CQRNN, validation period of CQRNN, training period of QR, validation period of QR, training periods of LR, MARS, KNN, and validation periods of LR, MARS, KNN, The 50% predictive interval is the area between the 25th and 75th percentile models. The 95% predictive interval is the area between the 2.5th and 97.5th percentile models

Figure 17 and Figure 18 compare the CQRNN model with the SFPUC's weekly regression model, respectively, by the time series and the scatter plot. Figure 17 and Figure 18 show that the SFPUC's weekly regression model cannot capture peaks caused by the rain-on-snow event in 1997, while CQRNN can reach these turbidity spikes. In addition, according to Figure 18, it is found that the scatter plot from CQRNN follows the 1-1 line better than the SFPUC's weekly regression turbidity model at the O'Shaughnessy Dam. Additionally, CQRNN has a smaller mean absolute error than QR. Consequently, it is concluded that CQRNN outperforms the SFPUC regression model.

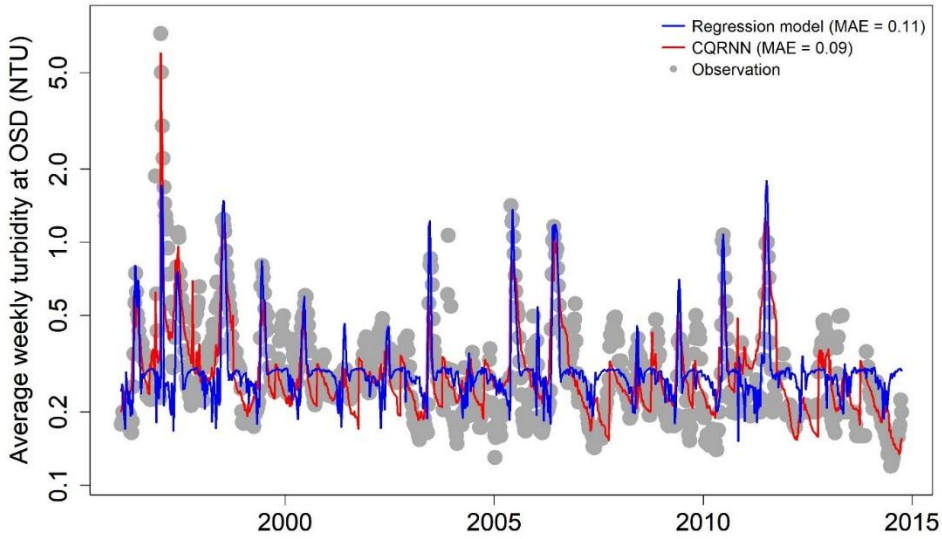

Figure 17 Compare CQRNN with the SFPUC's weekly regression model by the scatter plot for turbidity at the O'Shaughnessy Dam (OSD). MAE is the mean absolute error between predicted and observed turbidity values

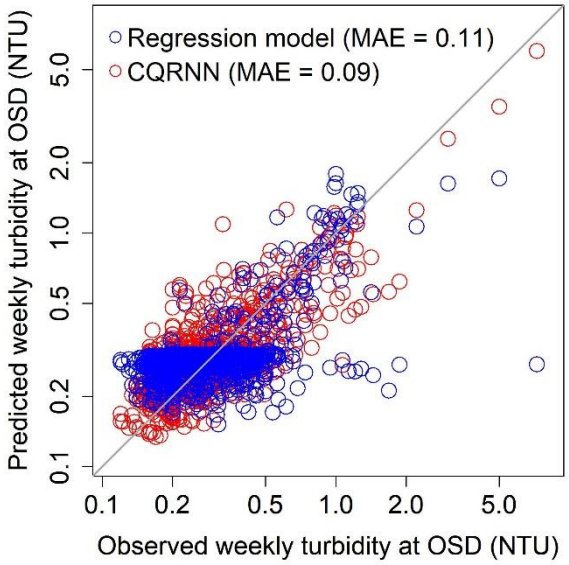

Figure 18 Compare CQRNN with the SFPUC's weekly regression model by the time series plot for turbidity at the O'Shaughnessy Dam (OSD). MAE is the mean absolute error between predicted and observed turbidity values

4.2.2 Turbidity at the Tesla Portal

The historical TOC data is less variable and does not include any recorded extreme events, whereas the actual turbidity data contained the turbidity extreme on January 3, 1997. Thus, we compare the abilities to simulate extreme events from five models (i.e., CQRNN, QR, LR, MARS, and KNN) based on the turbidity event on January 3, 1997. This comparison is shown in Figure 19.

On January 2, 1997, a massive inflow caused by a rain-on-snow event entered the Hetch Hetchy Reservoir (see the middle right plot of Figure 19). This discharge peak caused a spike in turbidity level (6.77 NTU) on January 3, 1997 (see the top right graph of Figure 19). In this situation, reservoir operators reduced the SJPL flow on the same date and shut down these water supply pipelines completely one day after that (see the bottom right plot of Figure 19). Only CQRNN can capture the turbidity peak in 1997, although all models use the 1997 event in their calibrations. Even the 97.5^{th} percentile prediction of QR cannot simulate precisely the turbidity spike in 1997 (see the middle left plot of Figure 19). In brief, CQRNN outperforms QR, LR, MARS, and KNN in predicting the 1997 turbidity extreme.

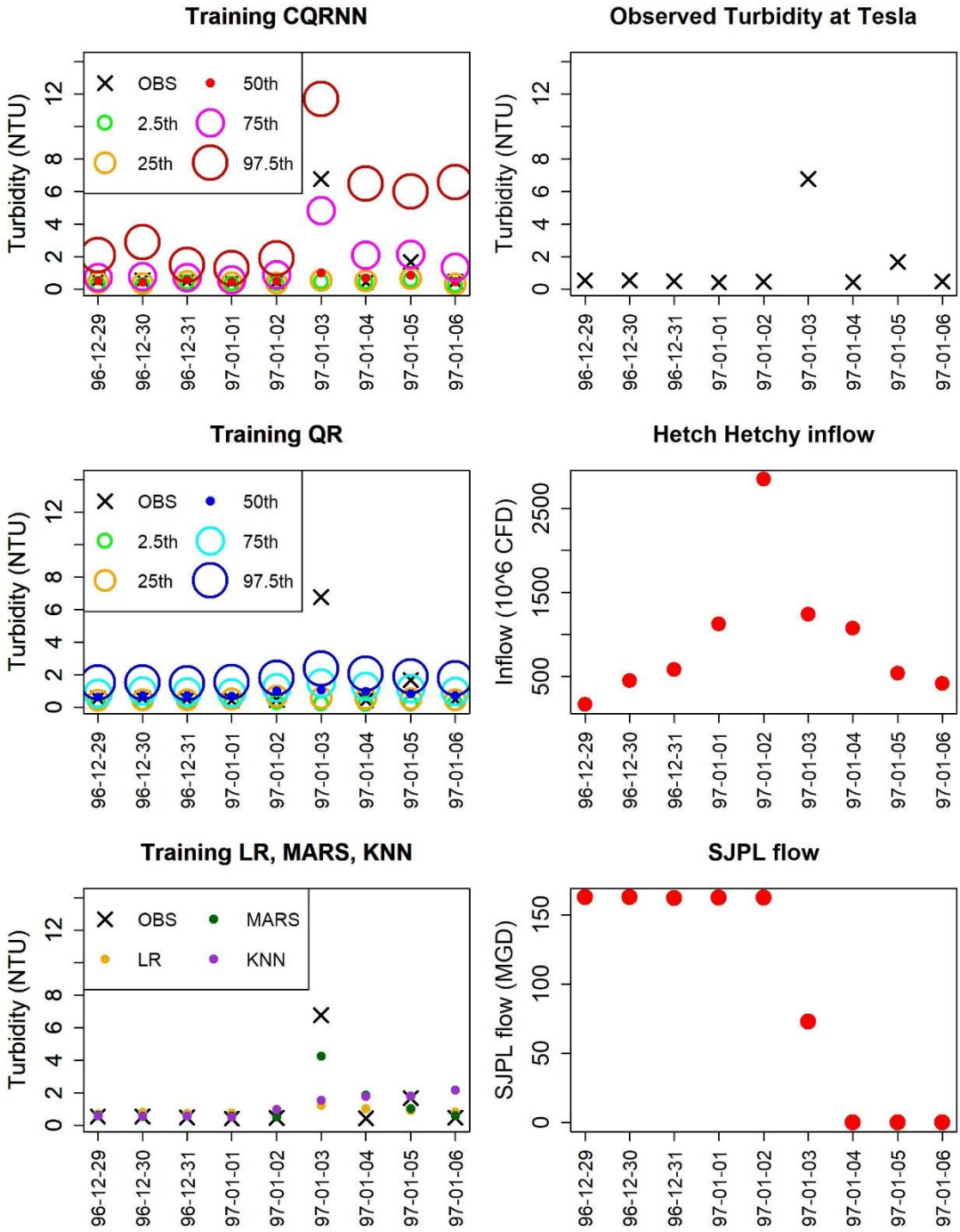

Figure 19 Compare the performance of CQRNN (Composite Quantile Regression Neural Network) with QR (Quantile Regression), LR (Linear Regression), MARS (Multivariate Adaptive Regression Splines), KNN (K-Nearest Neighbors) in modeling the turbidity peak at the Tesla Portal on January 3, 1997. The comparison is illustrated by simulations of 9 days (12/29/1996 - 01/06/1997) around the turbidity peak event at the Tesla Portal on January 3, 1997 in CQRNN (Composite Quantile Regression Neural Network), QR (Quantile Regression), LR (Linear Regression), MARS (Multivariate Adaptive Regression Spline), KNN (K-Nearest Neighbors). Left figures from top to bottom respectively illustrate CQRNN, QR, and three methods from LR, MARS, KNN. In top left and middle left figures, "2.5th", "25th", "50th", "75th", "97.5th" are respectively 2.5th, 25th, 50th, 75th, 97.5th percentile predictions. Right figures from top to bottom respectively show the observed turbidity levels (OBS), the Hetch Hetchy inflow, and the San Joaquin Pipeline (SJPL) flow during these 9 days

Figure 20 reveals the performances of five methods (i.e., CQRNN, QR, LR, MARS, and KNN) in simulating high turbidity levels at the Tesla Portal. Specifically, in the left plot of Figure 20, CDFs from five techniques are shown for turbidity greater than 0.6 NTU and for the cumulative probability above 0.95. It is found that only the 97.5^{th} percentile prediction from CQRNN captures turbidity levels above 2 NTU (see the left plot of Figure 20). In addition, the CDF of median estimation from CQRNN is closer to the observed CDF than QR, LR, and MARS. The CDF of MARS does not appear in the left plot of Figure 20 because it is outside of axes' limits, implying the poor performance. The high loss and AE also prove this unsatisfactory predictive performance (see

Table 5). Likewise, the CDF of median prediction from KNN follows the observed CDF well, but this model cannot re-produce turbidity values higher than 2 NTU. In brief, only CQRNN has a satisfactory predictive performance in modeling the mean and extremes (i.e., higher than 2 NTU) for turbidity at the Tesla Portal.

In the right plot of Figure 20, the quantile-quantile graph compares predicted and observed cumulative probabilities of turbidity at the Tesla Portal above 0.65. It is noted that high quantile estimations (i.e., 97.5th percentile ones) of CQRNN and QR follow the 1-1 line (the black dashed line in Figure B) better than median predictions from LR, MARS, KNN. Thus, performances of high quantile estimations are superior to median estimations for high turbidity levels at the Tesla Portal. Furthermore, CQRNN outperforms QR for extremes whose cumulative probabilities greater than 0.9 since CQRNN follows the 1-1 line more tightly than QR (see the right plot of Figure 20). In summary, CQRNN has a better performance than other conventional techniques such as QR, LR, MARS, and KNN in simulating turbidity peaks at the Tesla Portal in HHRWS.

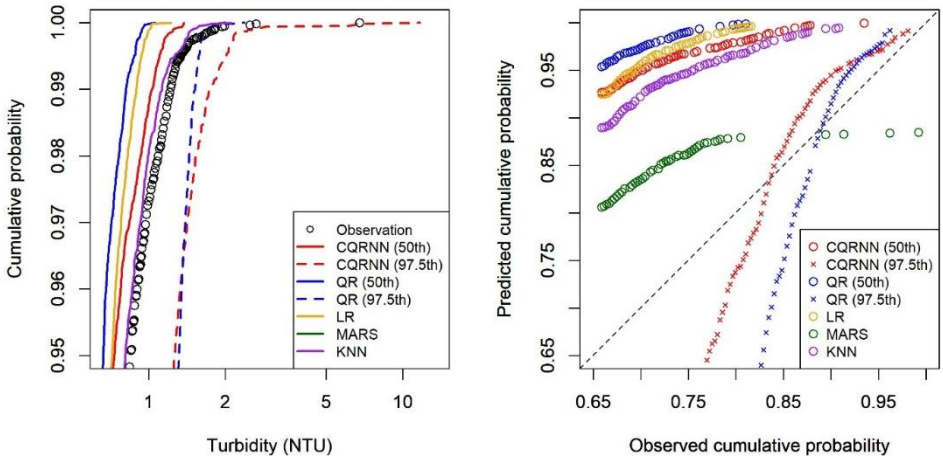

Figure 20 Compare the performance of CQRNN (Composite Quantile Regression Neural Network) with QR (Quantile Regression), LR (Linear Regression), MARS (Multivariate Adaptive Regression Splines), KNN (K-Nearest Neighbors) in modeling high turbidity values at the Tesla Portal. In the left figure, the comparison is illustrated by plotting cumulative density functions (CDFs) of high turbidity levels at the Tesla Portal (i.e., approximately turbidity is greater than 0.6 NTU and the cumulative probability is larger than 0.95) for median ("50th") and 97.5th percentile ("97.5th") predictions of CQRNN and QR, for median estimations of LR, MARS, KNN. It is noted that MARS does not appear in the left figure because its CDF exceeds the limits of horizontal and vertical axes, which reveals the poor performance in the simulation of turbidity peaks at the Tesla Portal. In the right figure, the comparison is shown by plotting predicted versus observed cumulative probabilities of high turbidity levels at the Tesla Portal (i.e., cumulative probabilities are higher than 0.65)

To emphasize the comparison by losses and AEs in

Table 5, Figure 21 and Figure 22 respectively show the time series and the CDF of turbidity at the Tesla Portal for five models (CQRNN, QR, LR, MARS, and KNN). We observe that CQRNN and QR models estimate the conditional median with a 95% central prediction interval generated by five percentile estimations (2.5^{th}, 25^{th}, 50^{th}, 75^{th}, and 97.5^{th}). Thus, CQRNN and QR can simulate very high and low water quality levels. Points outside the 95% central prediction interval of CQRNN and QR are fewer than uncaptured points in LR, MARS, and KNN techniques (see Figure 21 and Figure 22). The prediction range is not available in LR, MARS, and KNN because they do not provide multiple quantile estimations.

CQRNN outperforms QR based on losses and AEs in

Table 5. Furthermore, in Figure 21, QR has a poorer performance than CQRNN at the validation period after 2014 (see the middle right plot of Figure 21). Figure 22 provides a better view of predictive performances between CQRNN and QR. In Figure 22, the grey area of QR, representing the 95% prediction interval, is wider than the one of CQRNN. This finding implies QR has a higher degree of underestimation and overestimation respectively from 2.5^{th} and 97.5^{th} percentile estimations. In addition, QR encounters the issue of crossing quantiles, which produces nonphysical results. CQRNN uses simultaneous calibrations for all quantiles, which avoids this problem and generates more physical estimations. In brief, CQRNN has a better overall performance than QR, LR, MARS, and KNN in modeling median and other data quantiles.

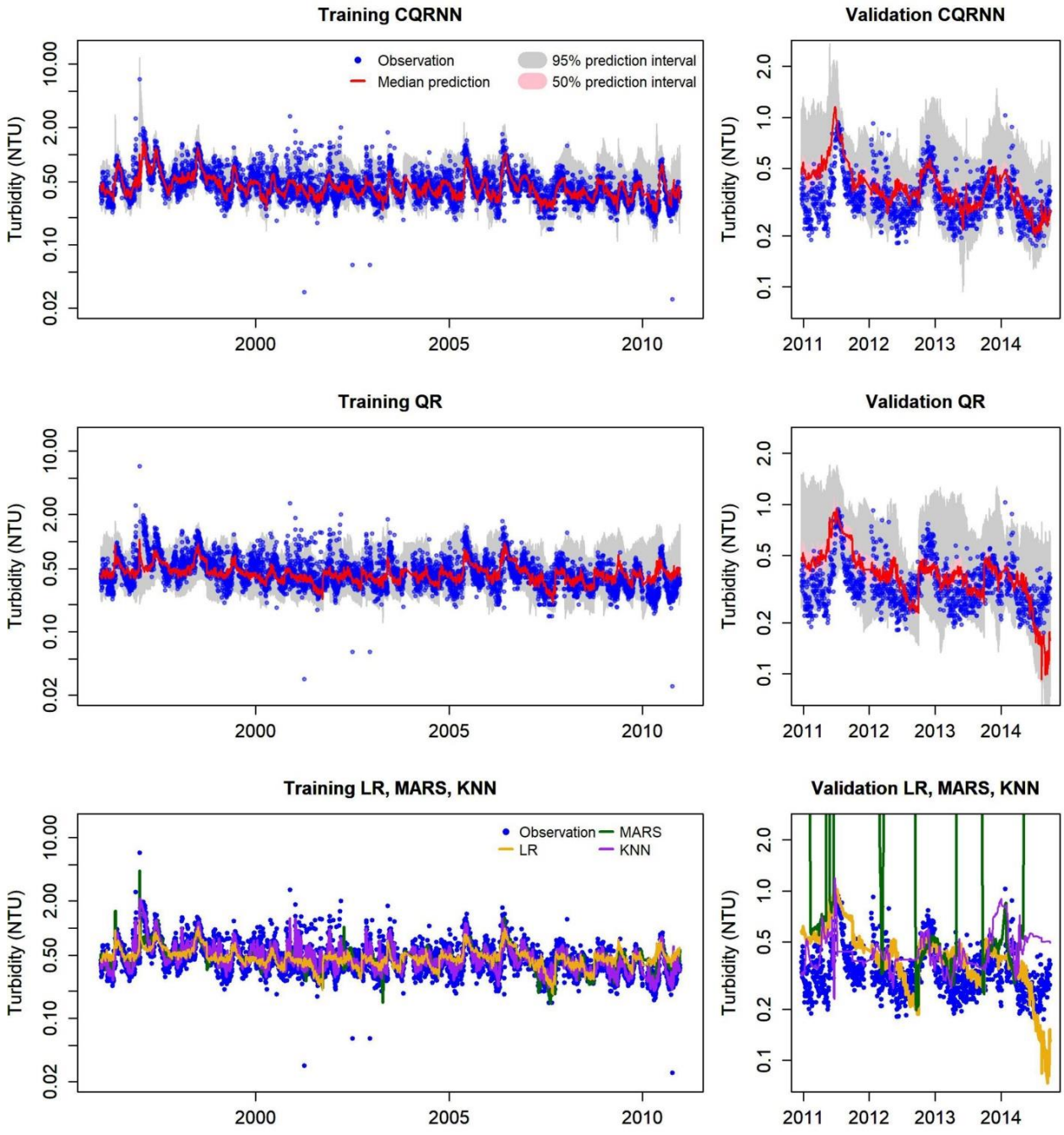

Figure 21 Predicted and observed time series of turbidity at the Tesla Portal in training and validation periods for five methods: CQRNN (Composite Quantile Regression Neural Network), QR (Quantile Regression), LR (Linear Regression), MARS (Multivariate Adaptive Regression Splines), KNN (K-Nearest Neighbors). Top left, top right, middle left, middle right, bottom left and bottom right respectively illustrate training period of CQRNN, validation period of CQRNN, training period of QR, validation period of QR, training periods of LR, MARS, KNN, and validation periods of LR, MARS, KNN, The 50% predictive interval is the area between the 25th and 75th percentile models. The 95% predictive interval is the area between the 2.5th and 97.5th percentile models

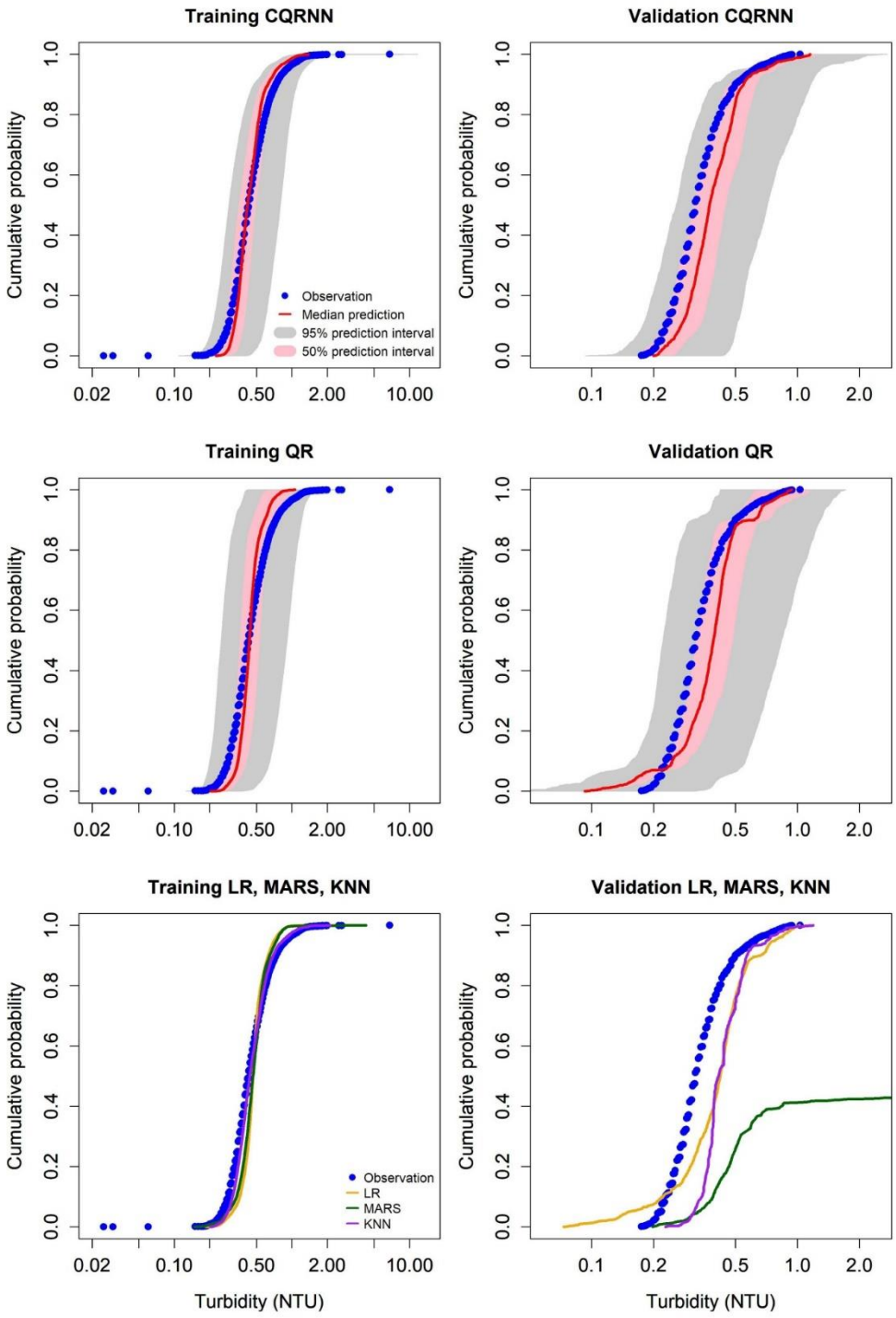

Figure 22 Predicted and observed cumulative density functions of turbidity at the Tesla Portal in training and validation periods for five methods: CQRNN (Composite Quantile Regression Neural Network), QR (Quantile Regression), LR (Linear Regression), MARS (Multivariate Adaptive Regression Splines), KNN (K-Nearest Neighbors). Top left, top right, middle left, middle right, bottom left and bottom right respectively illustrate training period of CQRNN, validation period of CQRNN, training period of QR, validation period of QR, training periods of LR, MARS, KNN, and validation periods of LR, MARS, KNN, The 50% predictive interval is the area between the 25th and 75th percentile models. The 95% predictive interval is the area between the 2.5th and 97.5th percentile models

4.2.3 TOC at the Tesla Portal

Figure 23 and Figure 24 illustrate the comparison among five methods (CQRNN, QR, LR, MARS, and KNN) respectively by the time series and the CDF for TOC prediction at the Tesla Portal. Based on

Table 5, CQRNN generally outperforms other techniques in losses and AEs. We note that all methods can correctly represent the seasonal variation of the TOC across the training dataset. However, CQRNN and QR can capture more TOC high values than LR, MARS, and KNN based on their 97.5^{th} percentile estimations (see Figure 23 and Figure 24).

The grey area of QR, illustrating the 95% predictive range, is larger than CQRNN (see Figure 23 and Figure 24). Therefore, QR seems to underestimate and overestimate TOC levels more than CQRNN. Furthermore, according to

Table 5, CQRNN outperforms QR based on losses and AEs. In conclusion, CQRNN and QR both perform similarly. Additionally, CQRNN is slightly better (but it should not be significant). The slight difference between CQRNN and QR could result from the fact that TOC data at the Tesla Portal does not have extremes (the rain-on-snow event in 1997) like turbidity at the Tesla Portal and at the O'Shaughnessy Dam. Therefore, to simulate a less variable dataset like TOC at the Tesla Portal, a linear-regression-based approach (i.e., QR) can still perform well.

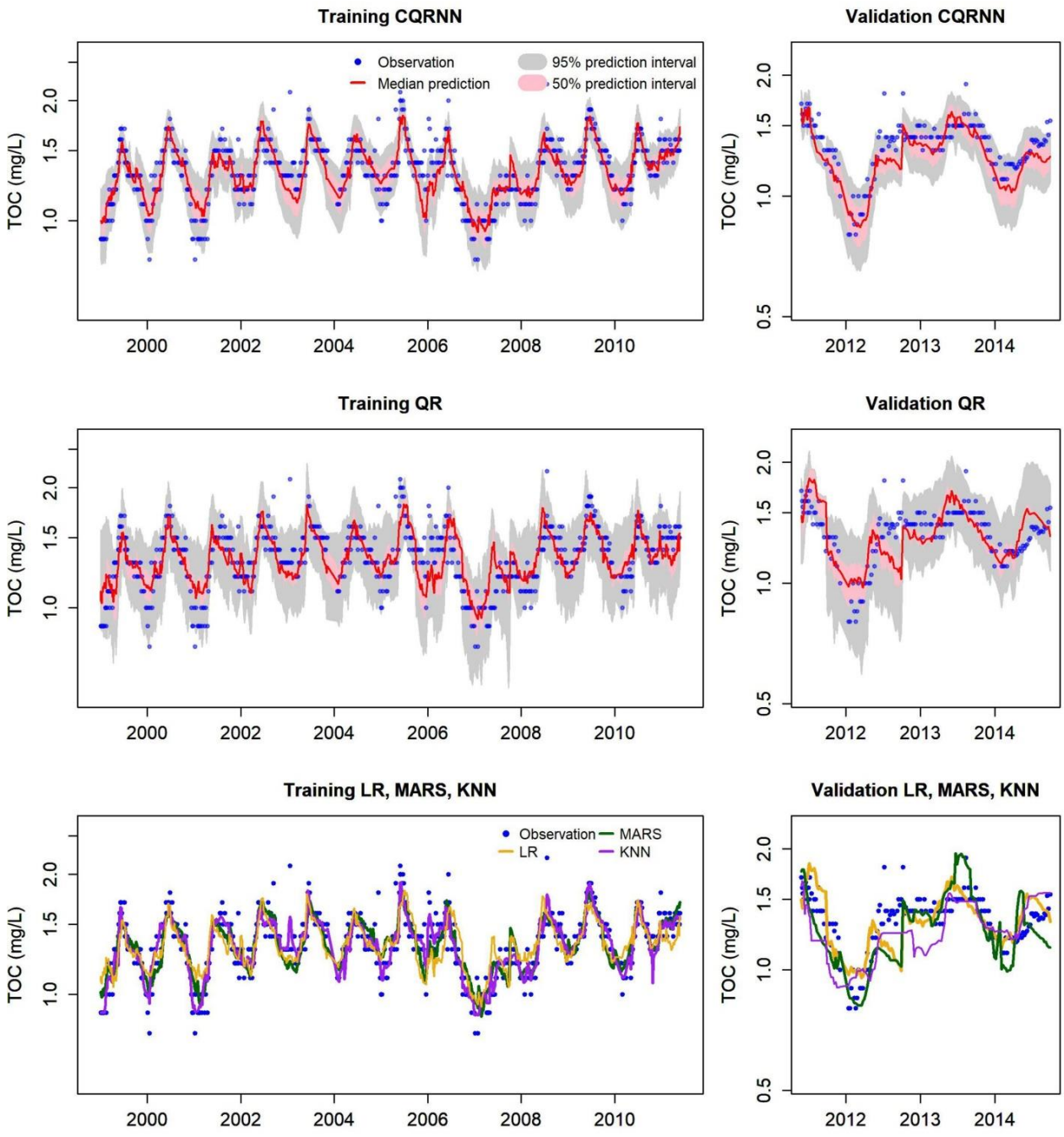

Figure 23 Predicted and observed time series of TOC at the Tesla Portal in training and validation periods for five methods: CQRNN (Composite Quantile Regression Neural Network), QR (Quantile Regression), LR (Linear Regression), MARS (Multivariate Adaptive Regression Splines), KNN (K-Nearest Neighbors). Top left, top right, middle left, middle right, bottom left and bottom right respectively illustrate training period of CQRNN, validation period of CQRNN, training period of QR, validation period of QR, training periods of LR, MARS, KNN, and validation periods of LR, MARS, KNN, The 50% predictive interval is the area between the 25th and 75th percentile models. The 95% predictive interval is the area between the 2.5th and 97.5th percentile models

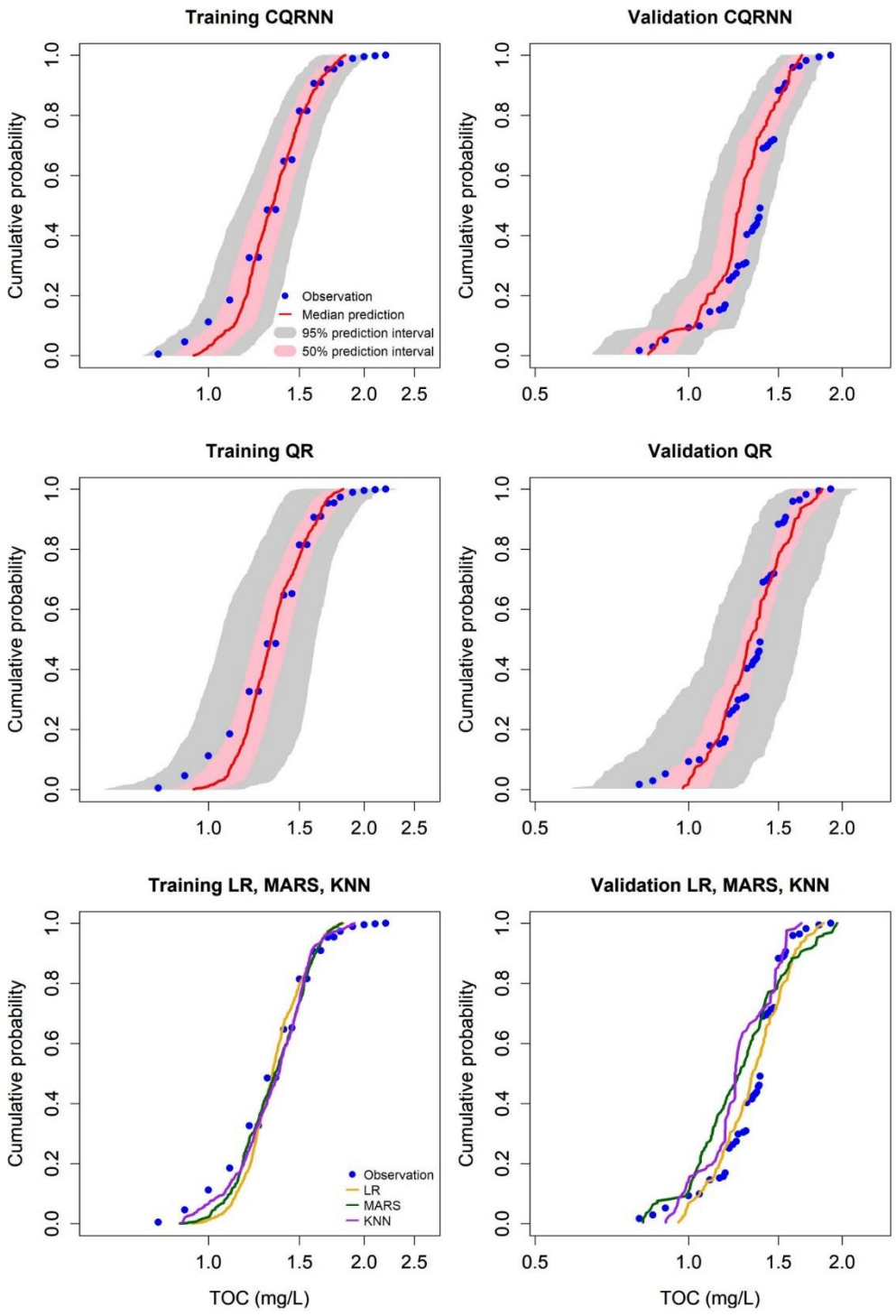

Figure 24 Predicted and observed cumulative density functions of TOC at the Tesla Portal in training and validation periods for five methods: CQRNN (Composite Quantile Regression Neural Network), QR (Quantile Regression), LR (Linear Regression), MARS (Multivariate Adaptive Regression Splines), KNN (K-Nearest Neighbors). Top left, top right, middle left, middle right, bottom left and bottom right respectively illustrate training period of CQRNN, validation period of CQRNN, training period of QR, validation period of QR, training periods of LR, MARS, KNN, and validation periods of LR, MARS, KNN, The 50% predictive interval is the area between the 25th and 75th percentile models. The 95% predictive interval is the area between the 2.5th and 97.5th percentile models

The developed CQRNN prediction model for TOC at the Tesla Portal has also been compared with the regression model currently used by SFPUC. This model is based on a linear regression model and conducted in an annual time step (Rob Clark 2017):

$Water\ year\ average\ Tesla\ TOC\ [mg/L] = 1.3103 + (0.03476 \times Z_1) - (0.02807 \times Z_2)$ (Rob Clark 2017)

Z_1: Current water-year inflow (100K acre-feet/year)

Z_2: Prior water-year inflow (100K acre-feet/year)

Figure 25 and Figure 26 show CQRNN has a better predictive performance in simulating the water-year TOC average than the SFPUC's linear regression. In particular, CQRNN follows the variability of observed TOC better than the linear regression. Moreover, it has a mean absolute error lower than the linear regression's one.

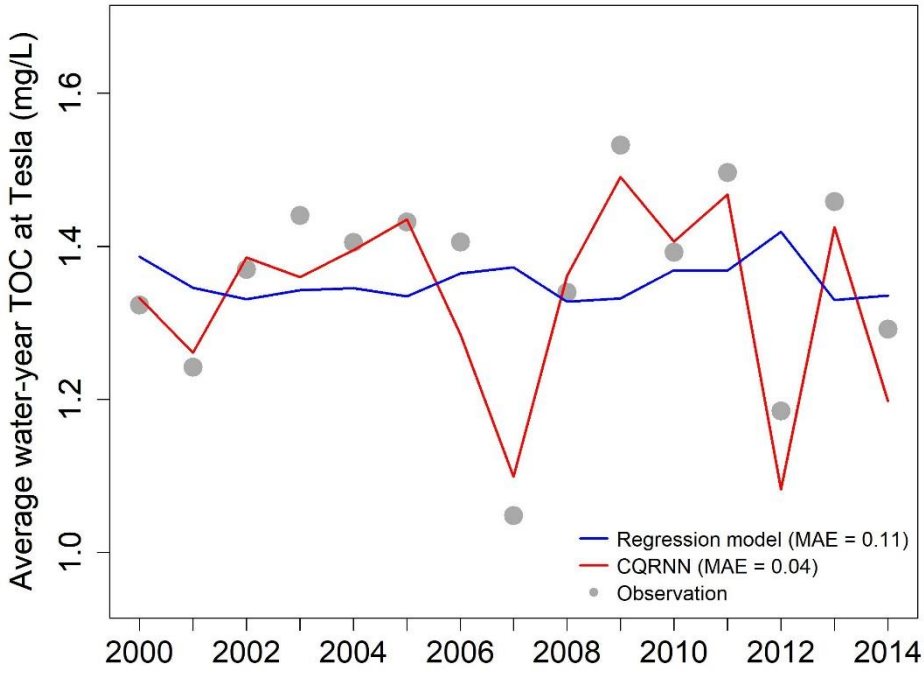

Figure 25 Compare CQRNN with the SFPUC's annual regression model by the time series plot for TOC at the Tesla Portal. MAE is the mean absolute error between predicted and observed TOC values

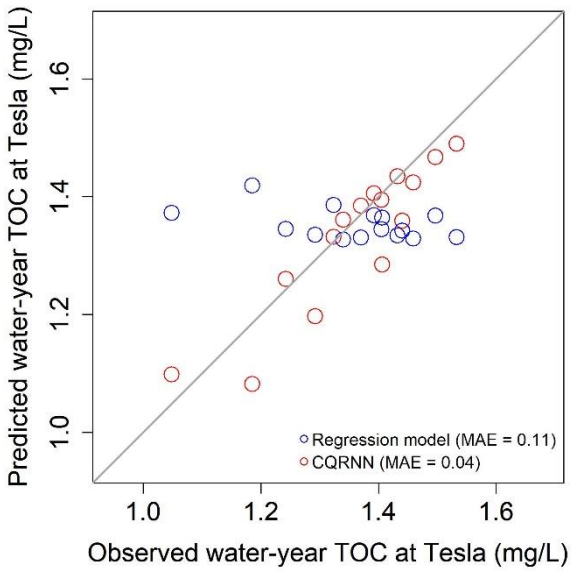

Figure 26 Compare CQRNN with the SFPUC's annual regression model by the scatter plot for TOC at the Tesla Portal. MAE is the mean absolute error between predicted and observed TOC values

In summary, CQRNN outperforms the four remaining methods (i.e., QR, LR, MARS, and KNN) based on losses and AEs. Additionally, it produces a median estimation with a 95% predictive range, assisting in predicting extraordinarily high and low values. All non-crossing conditional quantiles are close to true quantiles thanks to simultaneous simulations in CQRNN. Also, CQRNN can capture the turbidity extreme in 1997, while other methods cannot simulate this critical event. Thus, this model is helpful for water managers in modeling both means and peaks of water quality.

4.3 Contribution of the predictors to the CQRNN predictions

The Relative Importance metric (equation 7) is a way to assess the contribution from each predictor to the predicted value of considered water quality indicators. To complement the discussion of the relative importance (RI) metric, scatterplots between each predictor and turbidity, and TOC are given in Appendix 6.1.

For turbidity at the O'Shaughnessy Dam, the relative importance of predictors is shown in Figure 27. The primary predictors, which account for around 20% of relative importance, are dry days in the previous water year and discharges in the current water year. It emphasizes the importance of the annual water cycle for the turbidity prediction at the O'Shaughnessy Dam. Dry days of the prior water year refer to droughts, which can cause material accumulation in the watershed. Then, these materials are flushed out based on the hydrological condition during the current water year, generating suspended solids. The air temperature has the third-largest relative importance since it controls the snow melting stream since the Upcountry Watershed is a snow-dominated basin.

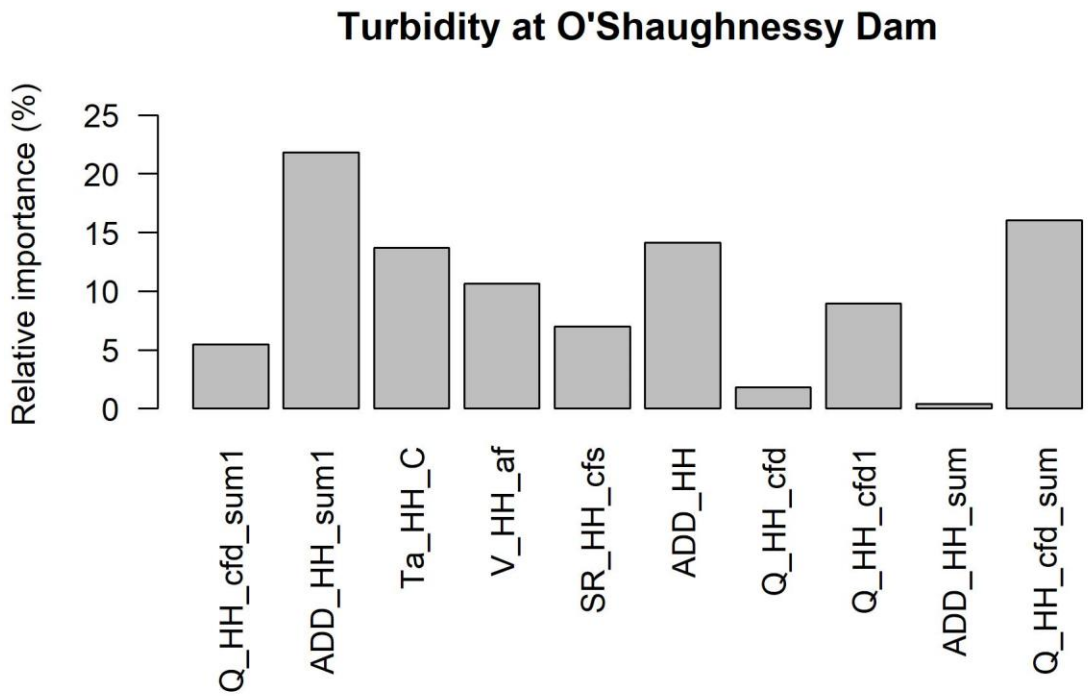

Figure 27 Predictor relative importance from CQRNN for turbidity at the O'Shaughnessy Dam model. Q_HH_cfd is the Hetch Hetchy inflow (Cubic feet per day, CFD). Ta_HH_C is Hetch Hetchy air temperature (Degree Celsius, °C). V_HH_af is the Hetch Hetchy storage (Acre-foot, AF). SR_HH_cfs is the Hetch Hetchy spill and release (Cubic foot per second, CFS). ADD_HH_sum1 is cumulative dry days in the previous water year (Days). Q_HH_cfd_sum1 is cumulative flows in the previous water year (Cubic foot, CF). ADD_HH_sum is cumulative dry days until the predicted day in the current water year (Days). Q_HH_cfd_sum is cumulative flows until the predicted day in the current water year (Cubic foot, CF). ADD_HH is the Hetch Hetchy antecedent dry days (days). Q_HH_cfd1 is the one-day-prior Hetch Hetchy inflow (Cubic foot, CFD)

Figure 28 illustrates the contribution from each predictor to the predicted turbidity at the Tesla Portal. Primary predictors for turbidity at the Tesla Portal, which account for more than 25% of relative importance, are discharges and dry days of the previous water year. The combination of droughts in material accumulation and stream flows in flushing triggers the turbidity level in the Hetch Hetchy Reservoir. This effect of the prior water year is significant for predicting turbidity because water is stored in the Hetch Hetchy Reservoir, a big lake and has a long-term hydrological memory. The third significant predictor is SJPL, whose relative importance is 15%. The rate of pipelines is crucial in simulating turbidity since higher water delivery brings more solids into pipelines and vice versa. Thus, the amount of water in SJPL governs suspended solids at the Tesla Portal.

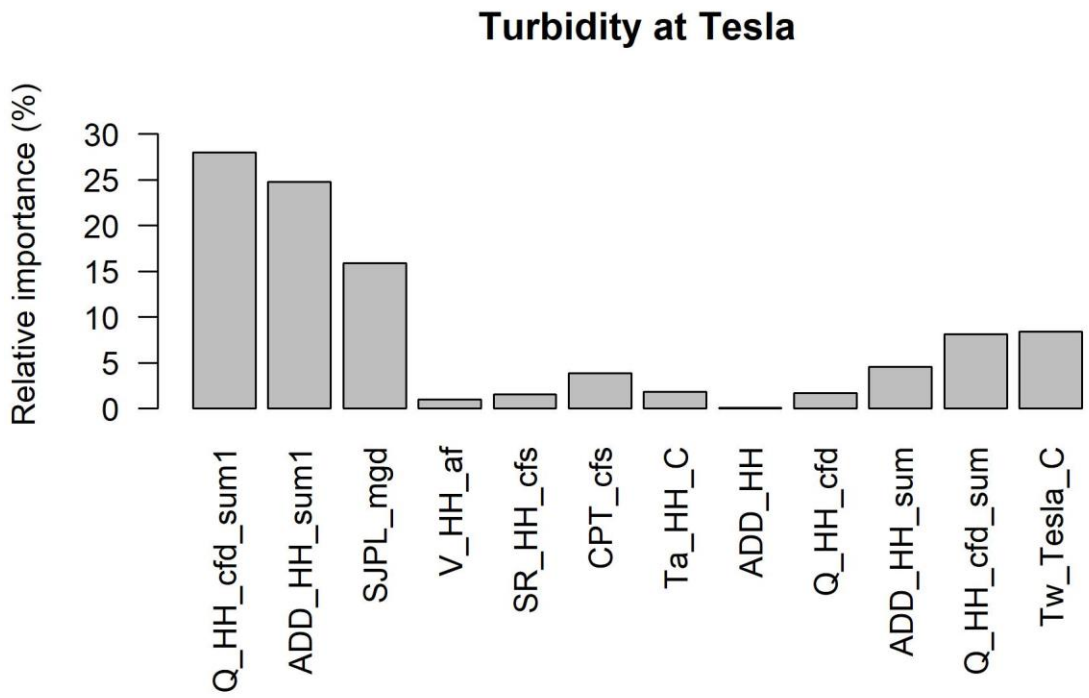

Figure 28 Predictor relative importance from CQRNN for turbidity at the Tesla Portal. Q_HH_cfd is the Hetch Hetchy inflow (Cubic feet per day, CFD). Ta_HH_C is Hetch Hetchy air temperature (Degree Celsius, °C). Tw_Tesla_C is the Tesla water temperature (Degree Celsius, °C). SJPL_mgd is the San Joaquin Pipeline flow (Millions of gallons per day, MGD). V_HH_af is the Hetch Hetchy storage (Acre-foot, AF). SR_HH_cfs is the Hetch Hetchy spill and release (Cubic foot per second, CFS). CPT_cfs is the Canyon Power Tunnel flow (Cubic foot per second, CFS). ADD_HH_sum1 is cumulative dry days in the previous water year (Days). Q_HH_cfd_sum1 is cumulative flows in the previous water year (Cubic foot, CF). ADD_HH_sum is cumulative dry days until the predicted day in the current water year (Days). Q_HH_cfd_sum is cumulative flows until the predicted day in the current water year (Cubic foot, CF). ADD_HH is the Hetch Hetchy antecedent dry days (days). Q_HH_cfd1 is the one-day-prior Hetch Hetchy inflow (Cubic foot, CFD)

Figure 29 shows the relative importance of predictors for modeling TOC at the Tesla Portal. The flow in the previous water year accounts for more than 25%, which is the most important in predicting TOC. The second vital predictors are the water temperature at the Tesla Portal and daily flow. They have similar relative importance values, approximately more than 15%. These findings are consistent with other papers, which also stated that discharges and temperatures control TOC. Clair and Ehrman (1998) and Ågren et al. (2007) found that the TOC tendency mirrors stream flows. Moreover, a higher temperature stimulates the ecosystem's biomass growth, contributing to the carbon pool (Evans, Monteith, and Cooper 2005; Ågren et al. 2007; Weyhenmeyer and Karlsson 2009; and Pagano, Bida, and Kenny 2014); therefore, TOC increases when temperature enhances

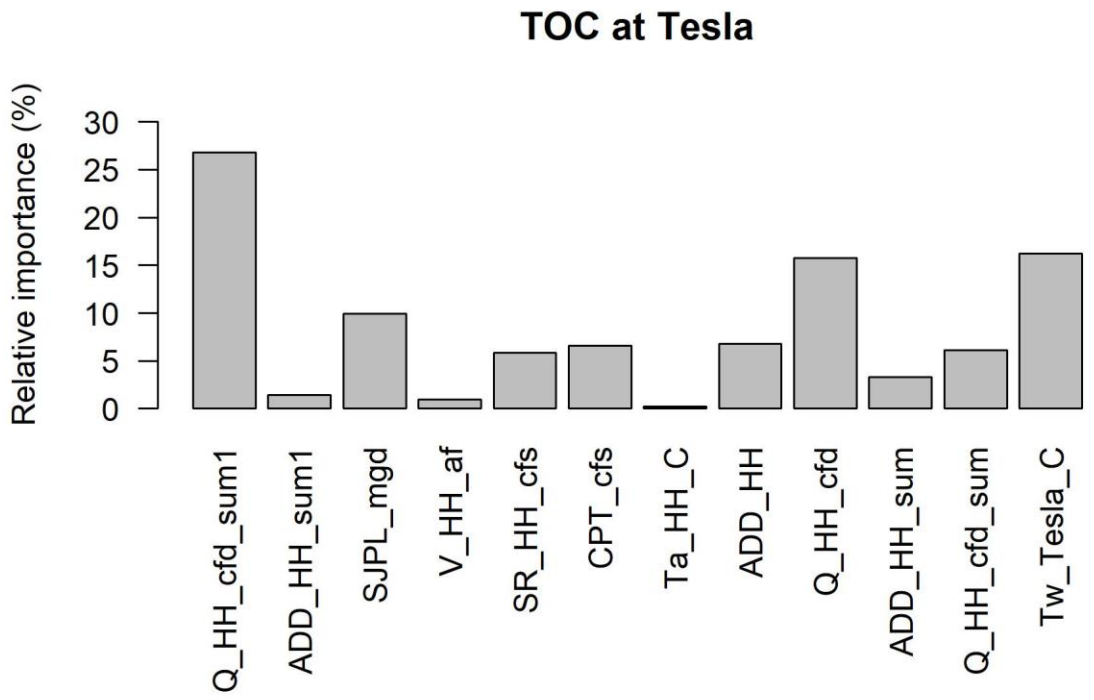

Figure 29 Predictor relative importance from CQRNN for total organic carbon (TOC) at the Tesla Portal. Q_HH_cfd is the Hetch Hetchy inflow (Cubic feet per day, CFD). Ta_HH_C is Hetch Hetchy air temperature (Degree Celsius, °C). Tw_Tesla_C is the Tesla water temperature (Degree Celsius, °C). SJPL_mgd is the San Joaquin Pipeline flow (Millions of gallons per day, MGD). V_HH_af is the Hetch Hetchy storage (Acre-foot, AF). SR_HH_cfs is the Hetch Hetchy spill and release (Cubic foot per second, CFS). CPT_cfs is the Canyon Power Tunnel flow (Cubic foot per second, CFS). ADD_HH_sum1 is cumulative dry days in the previous water year (Days). Q_HH_cfd_sum1 is cumulative flows in the previous water year (Cubic foot, CF). ADD_HH_sum is cumulative dry days until the predicted day in the current water year (Days). Q_HH_cfd_sum is cumulative flows until the predicted day in the current water year (Cubic foot, CF). ADD_HH is the Hetch Hetchy antecedent dry days (days). Q_HH_cfd1 is the one-day-prior Hetch Hetchy inflow (Cubic foot, CFD)

Although the RI metric is interesting as it directly evaluates the relative contribution for all predictors by analyzing the model parameters, it suffers from a major shortcoming. The RI metrics estimated from the model parameters indeed result from the training of the model. It is not possible to predict the relative importance of a specific predictor within a modified context. It is then necessary to remember that the relative importance metric discussed in this section for each predictor is relevant for the historical period. Still, it could not be for future uncertain conditions.

5 Conclusion

This technical report describes technically data-driven models that aim to predict turbidity at the O'Shaughnessy Dam, turbidity, and TOC at the Tesla Portal in the HHRWS. The Composite Quantile Regression Neural Network (CQRNN) produces low averaged quantile regression error functions over 2.5^{th}, 25^{th}, 50^{th}, 75^{th}, 97.5^{th} percentiles for turbidity and TOC models in all predicted locations. It is because turbidity and TOC are water quality indicators primarily being affected by physical factors included in the list of predictors used as input to the developed CQRNN. Furthermore, the CQRNN can capture nonlinear processes between predictors and predictands, which is the case for the formation of turbidity and TOC in the HHRWS. Five quantiles of the distribution of the considered water quality indicators have been considered in this work (2.5^{th}, 25^{th}, 50^{th}, $75^{th,}$ and 97.5^{th} percentiles), predicting an estimate of the conditional median and a 95% central prediction interval (Cannon 2011). Finally, this estimate helps models capture extreme water quality values crucial for water quality control.

In conclusion, models are successfully built for turbidity at the O'Shaughnessy Dam and for turbidity and TOC at the Tesla Portal, a primary compliant location in the HHRWS. In comparison with other models such as linear regression, quantile regression, multivariate adaptive regression spline, and K-nearest neighbors, CQRNN outperforms other techniques in terms of CQRNN's loss (averaged quantile regression error function over 2.5^{th}, 25^{th}, 50^{th}, 75^{th}, 97.5^{th} percentiles, described in equation 6). The next step for using this model is conducting the climate stress test on water quality in the HHRWS, particularly for turbidity at the O'Shaugnessy Dam, turbidity, and TOC at the Tesla portal, the primary compliant point of water quality in the HHRWS. This work is served as a step in the vulnerability assessment on the HHRWS.

6 Appendices:

6.1 Relation between predictand and predictors

6.1.1 Turbidity at O'Shaughnessy Dam

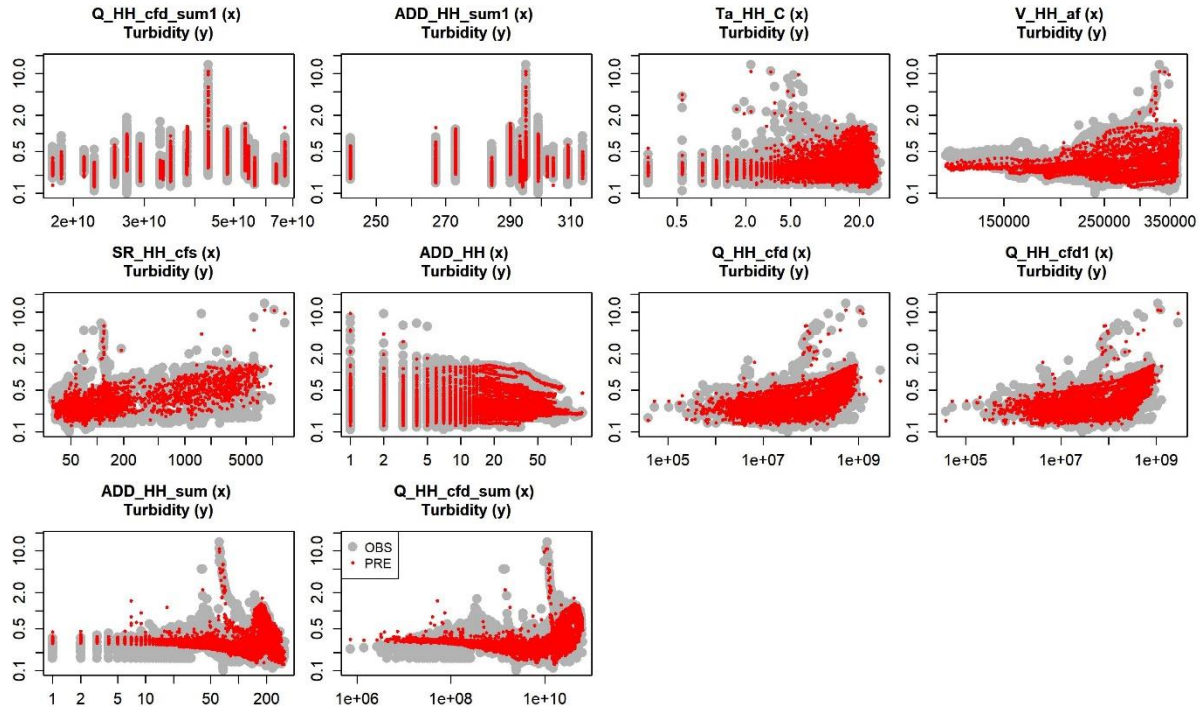

Figure 30 Relationships between turbidity at O'Shaughnessy Dam and predictors. OBS is the observed turbidity, PRE is the predicted turbidity. Short names of predictors are referred to Table 2

6.1.2 Turbidity at the Tesla Portal

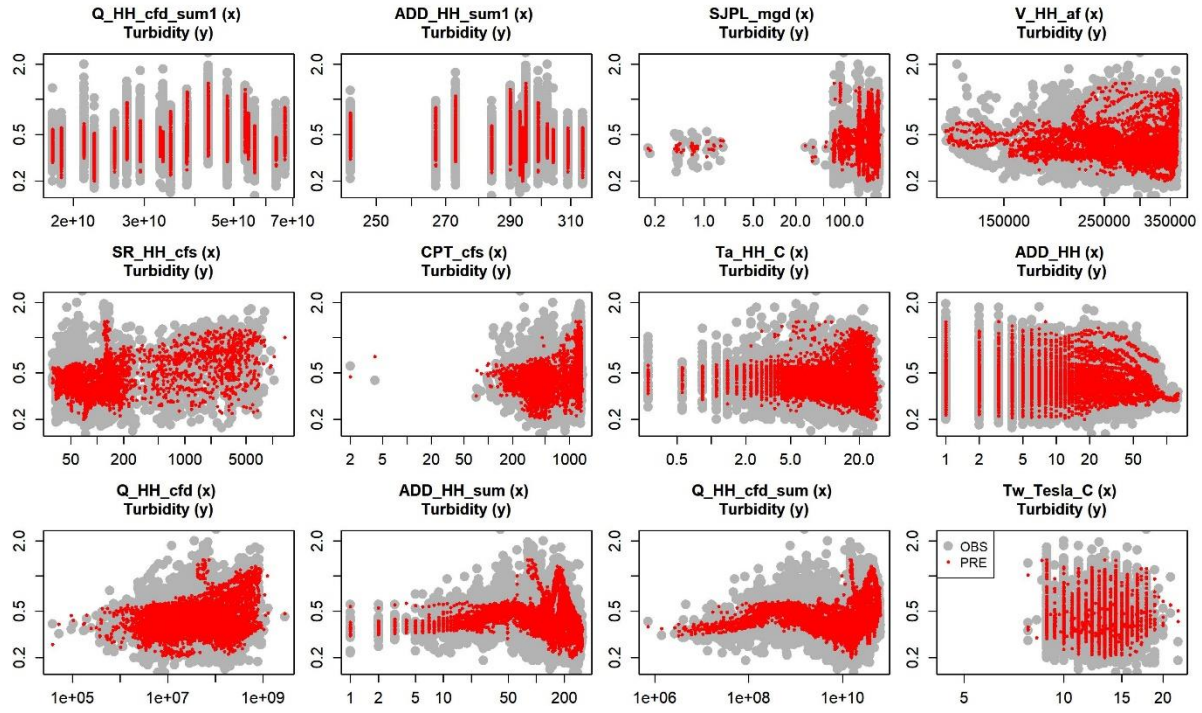

Figure 31 Relationships between turbidity at the Tesla Portal and predictors. OBS is the observed turbidity, PRE is the predicted turbidity. Short names of predictors are referred to Table 2

6.1.3 TOC at the Tesla Portal

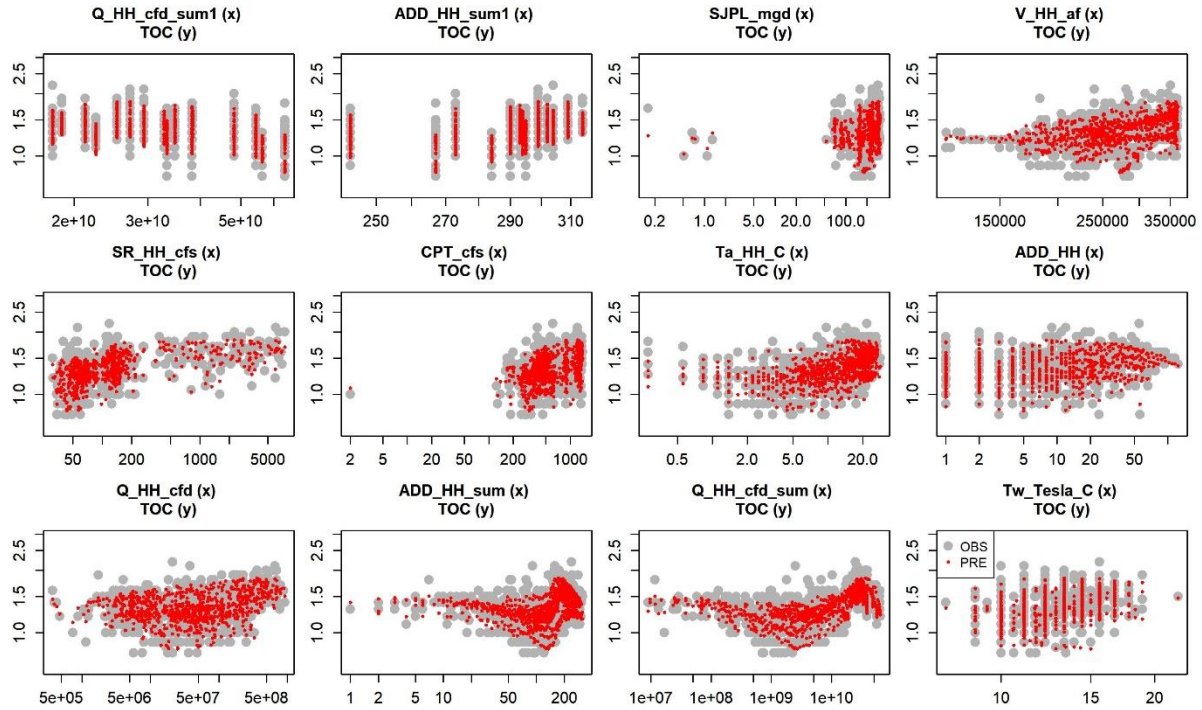

Figure 32 Relationships between TOC at the Tesla Portal and predictors. OBS is the observed TOC, PRE is the predicted TOC. Short names of predictors are referred to Table 2

6.2 Relation between air temperature and water temperature

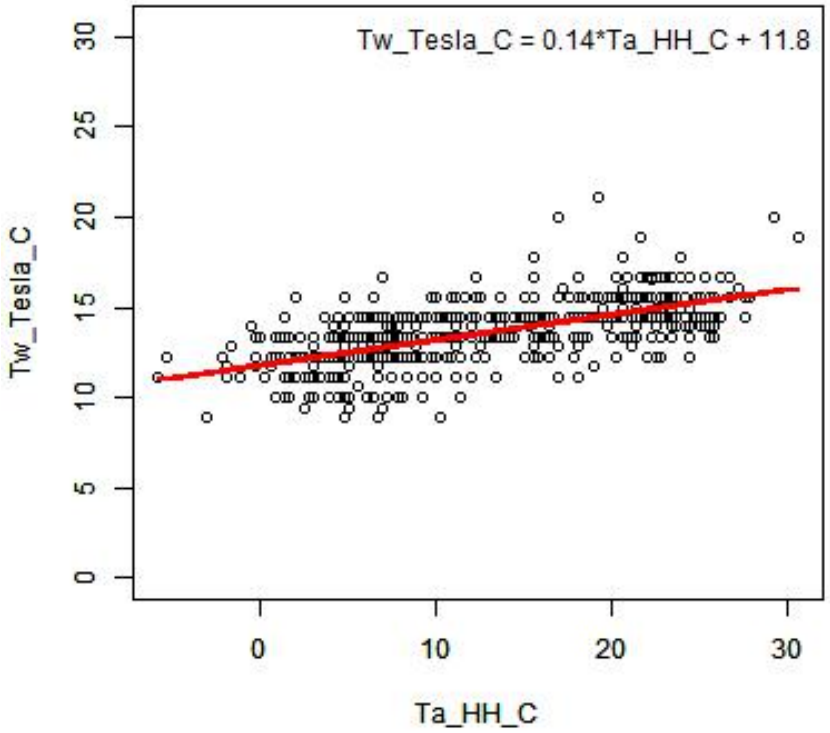

Figure 33 Linear relationship between water temperature at the Tesla Portal in degrees Celsius (Tw_Tesla_C) and air temperature at Hetch Hetchy in degrees Celsius (Ta_HH_C) with $R^2 = 0.46$

6.3 Recurrence interval of high turbidity event in 1997

We fit different distributions to annual maximum daily turbidity in the historical period 1997-2014 by L-moments (Hosking, Wallis, and Wood 1985), a linear combination of order statistics. This technique is used to define parameters of fitted distributions on a data sample. Figure 34 reveals the best-fitted distributions for annual maximum daily turbidity at Tesla, which are Generalized Extreme-value and Generalized Logistic. Figure 35 compares the exceedance probability curves between the empirical distribution and fitted distributions. The empirical distribution is defined by the Hazen method (Hosking and Wallis 1995). Generalized Extreme-value and Generalized Logistic are also the best-fitted curves based on Figure 35. By using Generalized Extreme-value and Generalized Logistic, we compute the exceedance probability (frequency of occurrence) of 0.015 and the return period of 67 years for the 1997 turbidity event (the highest turbidity level ever recorded at Tesla). We also calculate the exceedance probability and the return period for the second and the third highest turbidity levels ever recorded at Tesla, which are respectively 0.08 and 12 years (turbidity level of 2.65 NTU on 2000-11-22), and 0.09 and 11 years (turbidity level of 2.5 NTU on 1996-11-24).

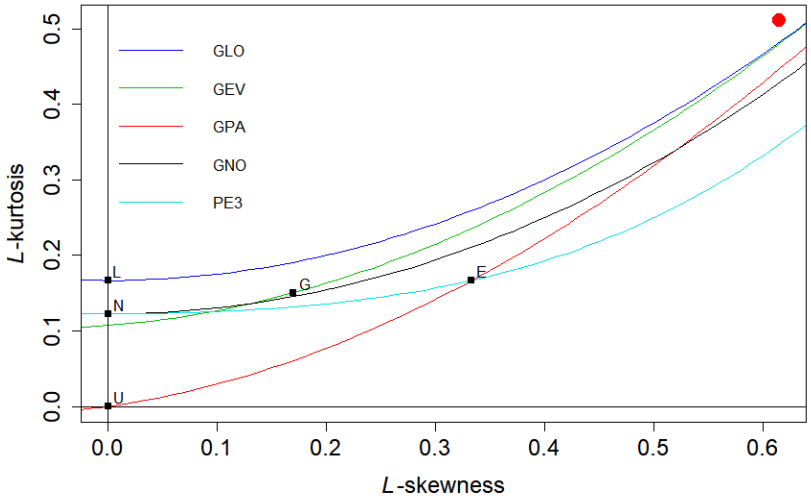

Figure 34 The L-moment ratio diagram of Generalized Logistics (GLO), Generalized Extreme-value (GEV), Generalized Pareto (GPA), Generalized Normal (GNO), Pearson type III (PE3), Logistic Distribution (L), Normal Distribution (N), Exponential Distribution (E), Gumbel Distribution (G), and Uniform Distribution (U). Lines are L-kurtosis and L-skewness parameters calculated from fitted distributions; while the red dot represents L-kurtosis and L-skewness parameters computed from the observed turbidity distribution

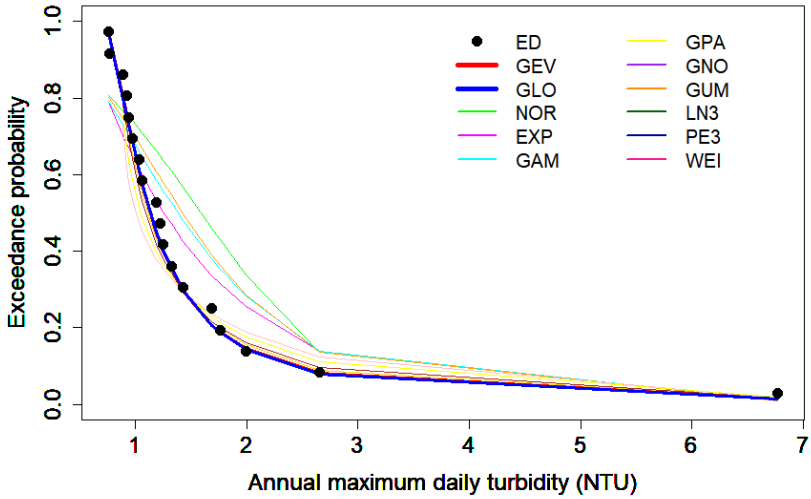

Figure 35 Compare exceedance probability curves of annual maximum daily turbidity between empirical distribution (ED) represented as black dots and lines which are fitted distributions including Generalized Extreme-value (GEV), Generalized Logistics (GLO), Normal Distribution (NOR), Exponential Distribution (EXP), Gamma Distribution (GAM), Generalized Pareto (GPA), Generalized Normal (GNO), Gumbel Distribution (GUM), Lognormal Distribution (LN3), Pearson type III (PE3), Weibull Distribution (WEI). The GLO line overlays the GEV line in the figure

7 References

Ågren, Anneli, Mats Jansson, Hans Ivarsson, Kevin Bishop, and Jan Seibert. 2007. "Seasonal and Runoff-Related Changes in Total Organic Carbon Concentrations in the River Öre, Northern Sweden." *Aquatic Sciences*. https://doi.org/10.1007/s00027-007-0943-9.

Bowden, Gavin J., Graeme C. Dandy, and Holger R. Maier. 2005. "Input Determination for Neural Network Models in Water Resources Applications. Part 1—Background and Methodology." *Journal of Hydrology* 301 (1): 75–92. https://doi.org/10.1016/j.jhydrol.2004.06.021.

Brown, Casey, Yonas Ghile, Mikaela Laverty, and Ke Li. 2012. "Decision Scaling: Linking Bottom-up Vulnerability Analysis with Climate Projections in the Water Sector." *Water Resources Research* 48 (9). https://doi.org/10.1029/2011WR011212.

Cannon, Alex J. 2011. "Quantile Regression Neural Networks: Implementation in R and Application to Precipitation Downscaling." *Computers & Geosciences* 37 (9): 1277–84. https://doi.org/10.1016/j.cageo.2010.07.005.

———. 2018. "Non-Crossing Nonlinear Regression Quantiles by Monotone Composite Quantile Regression Neural Network, with Application to Rainfall Extremes." *Stochastic Environmental Research and Risk Assessment* 32 (11): 3207–25. https://doi.org/10.1007/s00477-018-1573-6.

———. 2019. *Qrnn: Quantile Regression Neural Network* (version 2.0.5). https://CRAN.R-project.org/package=qrnn.

Chau, Kwok-wing. 2006. "A Review on Integration of Artificial Intelligence into Water Quality Modelling." *Marine Pollution Bulletin* 52 (7): 726–33. https://doi.org/10.1016/j.marpolbul.2006.04.003.

Clair, Thomas A., and James M. Ehrman. 1998. "Using Neural Networks to Assess the Influence of Changing Seasonal Climates in Modifying Discharge, Dissolved Organic Carbon, and Nitrogen Export in Eastern Canadian Rivers." *Water Resources Research* 34 (3): 447–55. https://doi.org/10.1029/97WR03472.

Costa, Cássia Monteiro da Silva Burigato, Izabel Rodrigues Leite, Aleska Kaufmann Almeida, and Isabel Kaufmann de Almeida. 2021. "Choosing an Appropriate Water Quality Model—a Review." *Environmental Monitoring and Assessment* 193 (1): 38. https://doi.org/10.1007/s10661-020-08786-1.

Cunningham, Pádraig, John Carney, and Saji Jacob. 2000. "Stability Problems with Artificial Neural Networks and the Ensemble Solution." *Artificial Intelligence in Medicine* 20 (3): 217–25. https://doi.org/10.1016/S0933-3657(00)00065-8.

Delpla, I., A. -V. Jung, E. Baures, M. Clement, and O. Thomas. 2009. "Impacts of Climate Change on Surface Water Quality in Relation to Drinking Water Production." *Environment International* 35 (8): 1225–33. https://doi.org/10.1016/j.envint.2009.07.001.

Dhillon, Gurbir Singh, and Shreeram Inamdar. 2013. "Extreme Storms and Changes in Particulate and Dissolved Organic Carbon in Runoff: Entering Uncharted Waters?" *Geophysical Research Letters* 40 (7): 1322–27. https://doi.org/10.1002/grl.50306.

Dreiseitl, Stephan, and Lucila Ohno-Machado. 2002. “Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review.” *Journal of Biomedical Informatics* 35 (5): 352–59. https://doi.org/10.1016/S1532-0464(03)00034-0.

Evans, C. D., D. T. Monteith, and D. M. Cooper. 2005. “Long-Term Increases in Surface Water Dissolved Organic Carbon: Observations, Possible Causes and Environmental Impacts.” *Environmental Pollution*, Recovery from acidificationin the UK: Evidence from 15 years of acid waters monitoring, 137 (1): 55–71. https://doi.org/10.1016/j.envpol.2004.12.031.

Francke, T., J. A. López-Tarazón, and B. Schröder. 2008. “Estimation of Suspended Sediment Concentration and Yield Using Linear Models, Random Forests and Quantile Regression Forests.” *Hydrological Processes* 22 (25): 4892–4904. https://doi.org/10.1002/hyp.7110.

Friedman, Jerome H. 1991. “Multivariate Adaptive Regression Splines.” *The Annals of Statistics* 19 (1): 1–67. https://doi.org/10.1214/aos/1176347963.

Futter, Martyn N., and Heleen A. de Wit. 2008. “Testing Seasonal and Long-Term Controls of Streamwater DOC Using Empirical and Process-Based Models.” *The Science of the Total Environment* 407 (1): 698–707. https://doi.org/10.1016/j.scitotenv.2008.10.002.

Gevrey, Muriel, Ioannis Dimopoulos, and Sovan Lek. 2003. “Review and Comparison of Methods to Study the Contribution of Variables in Artificial Neural Network Models.” *Ecological Modelling*, Modelling the structure of acquatic communities: concepts, methods and problems., 160 (3): 249–64. https://doi.org/10.1016/S0304-3800(02)00257-0.

Goh, A. T. C. 1995. “Back-Propagation Neural Networks for Modeling Complex Systems.” *Artificial Intelligence in Engineering* 9 (3): 143–51. https://doi.org/10.1016/0954-1810(94)00011-S.

Heddam, Salim, and Ozgur Kisi. 2018. “Modelling Daily Dissolved Oxygen Concentration Using Least Square Support Vector Machine, Multivariate Adaptive Regression Splines and M5 Model Tree.” *Journal of Hydrology* 559 (April): 499–509. https://doi.org/10.1016/j.jhydrol.2018.02.061.

Hosking, J. R. M., and J. R. Wallis. 1995. “A Comparison of Unbiased and Plotting-Position Estimators of L Moments.” *Water Resources Research* 31 (8): 2019–25. https://doi.org/10.1029/95WR01230.

Hosking, J. R. M., J. R. Wallis, and E. F. Wood. 1985. “Estimation of the Generalized Extreme-Value Distribution by the Method of Probability-Weighted Moments.” *Technometrics* 27 (3): 251–61. https://doi.org/10.1080/00401706.1985.10488049.

HRG LTVA. 2021. “Long Term Vulnerability Assessment and Adaptation Plan for the SFPUC Water Enterprise – Phase 1.” Hydrosystems Research Group, University of Massachusetts, Amherst, Amherst, Massachusetts.

HRG TR4. 2021. “Technical Report 4: San Francisco Water System Model.” Hydrosystems Research Group, University of Massachusetts, Amherst, Amherst, Massachusetts.

Kemp, Stanley J., Patricia Zaradic, and Frank Hansen. 2007. “An Approach for Determining Relative Input Parameter Importance and Significance in Artificial Neural Networks.” *Ecological Modelling* 204 (3): 326–34. https://doi.org/10.1016/j.ecolmodel.2007.01.009.

Khalil, B., T. B. M. J. Ouarda, and A. St-Hilaire. 2011. “Estimation of Water Quality Characteristics at Ungauged Sites Using Artificial Neural Networks and Canonical Correlation Analysis.” *Journal of Hydrology* 405 (3): 277–87. https://doi.org/10.1016/j.jhydrol.2011.05.024.

Koenker, Roger, Stephen Portnoy (Contributions to Censored QR code), Pin Tian Ng (Contributions to Sparse QR code), Achim Zeileis (Contributions to dynrq code essentially identical to his dynlm code), Philip Grosjean (Contributions to nlrq code), Cleve Moler (author of several linpack routines), and Brian D. Ripley (Initial (2001) R. port from S. (to my everlasting shame-- how could I. have been so slow to adopt R!) and for numerous other suggestions and useful advice). 2019. *Quantreg: Quantile Regression* (version 5.54). https://CRAN.R-project.org/package=quantreg.

Koenker, Roger W., and Gilbert Bassett. 1978. “Regression Quantiles.” *Econometrica* 46 (1): 33–50.

Lall, Upmanu, and Ashish Sharma. 1996. “A Nearest Neighbor Bootstrap For Resampling Hydrologic Time Series.” *Water Resources Research* 32 (3): 679–93. https://doi.org/10.1029/95WR02966.

Lee, Byoung-Hwa, and Miklas Scholz. 2006. “A Comparative Study: Prediction of Constructed Treatment Wetland Performance with k-Nearest Neighbors and Neural Networks.” *Water, Air, and Soil Pollution* 174 (1): 279–301. https://doi.org/10.1007/s11270-006-9113-2.

Leisch, S. original by Trevor Hastie & Robert Tibshirani Original R. port by Friedrich, and Kurt Hornik and Brian D. Ripley. 2017. *Mda: Mixture and Flexible Discriminant Analysis* (version 0.4-10). https://CRAN.R-project.org/package=mda.

Maier, Holger R., and Graeme C. Dandy. 2000. “Neural Networks for the Prediction and Forecasting of Water Resources Variables: A Review of Modelling Issues and Applications.” *Environmental Modelling & Software* 15 (1): 101–24. https://doi.org/10.1016/S1364-8152(99)00007-9.

Mukundan, Rajith, Donald C. Pierson, Lucien Wang, Adao H. Matonse, Nihar R. Samal, Mark S. Zion, and Elliot M. Schneiderman. 2013. “Effect of Projected Changes in Winter Streamflow on Stream Turbidity, Esopus Creek Watershed in New York, USA.” *Hydrological Processes* 27 (21): 3014–23. https://doi.org/10.1002/hyp.9824.

Najah, A., A. El-Shafie, O. A. Karim, and Amr H. El-Shafie. 2013. “Application of Artificial Neural Networks for Water Quality Prediction.” *Neural Computing and Applications* 22 (1): 187–201. https://doi.org/10.1007/s00521-012-0940-3.

Null, Sarah E., and Jay R. Lund. 2006. “Reassembling Hetch Hetchy: Water Supply Without O’shaughnessy Dam1.” *JAWRA Journal of the American Water Resources Association* 42 (2): 395–408. https://doi.org/10.1111/j.1752-1688.2006.tb03846.x.

Oña, Juan de, and Concepción Garrido. 2014. “Extracting the Contribution of Independent Variables in Neural Network Models: A New Approach to Handle Instability.” *Neural Computing and Applications* 25 (3): 859–69. https://doi.org/10.1007/s00521-014-1573-5.

Pagano, Todd, Morgan Bida, and Jonathan E. Kenny. 2014. “Trends in Levels of Allochthonous Dissolved Organic Carbon in Natural Water: A Review of Potential Mechanisms under a Changing Climate.” *Water* 6 (10): 2862–97. https://doi.org/10.3390/w6102862.

Pentoś, Katarzyna. 2016. “The Methods of Extracting the Contribution of Variables in Artificial Neural Network Models – Comparison of Inherent Instability.” *Computers and Electronics in Agriculture* 127 (September): 141–46. https://doi.org/10.1016/j.compag.2016.06.010.

Rob Clark. 2017. “Wet/Dry Year Impacts on Hetch Hetchy Source Water Quality.” November 8.

Samal, Nihar R., Adao H. Matonse, Rajith Mukundan, Mark S. Zion, Donald C. Pierson, Rakesh K. Gelda, and Elliot M. Schneiderman. 2013a. “Modelling Potential Effects of Climate Change on Winter Turbidity Loading in the Ashokan Reservoir, NY.” *Hydrological Processes* 27 (21): 3061–74. https://doi.org/10.1002/hyp.9910.

———. 2013b. “Modelling Potential Effects of Climate Change on Winter Turbidity Loading in the Ashokan Reservoir, NY.” *Hydrological Processes* 27 (21): 3061–74. https://doi.org/10.1002/hyp.9910.

Shu, C., and T. B. M. J. Ouarda. 2007. “Flood Frequency Analysis at Ungauged Sites Using Artificial Neural Networks in Canonical Correlation Analysis Physiographic Space.” *Water Resources Research* 43 (7). https://doi.org/10.1029/2006WR005142.

Sinshaw, Tadesse A., Cristiane Q. Surbeck, Hakan Yasarer, and Yacoub Najjar. 2019. “Artificial Neural Network for Prediction of Total Nitrogen and Phosphorus in US Lakes.” *Journal of Environmental Engineering* 145 (6): 04019032. https://doi.org/10.1061/(ASCE)EE.1943-7870.0001528.

Taylor, James W. 2000. “A Quantile Regression Neural Network Approach to Estimating the Conditional Density of Multiperiod Returns.” *Journal of Forecasting* 19 (4): 299–311. https://doi.org/10.1002/1099-131X(200007)19:4<299::AID-FOR775>3.0.CO;2-V.

Tiyasha, Tran Minh Tung, and Zaher Mundher Yaseen. 2020. “A Survey on River Water Quality Modelling Using Artificial Intelligence Models: 2000–2020.” *Journal of Hydrology* 585 (June): 124670. https://doi.org/10.1016/j.jhydrol.2020.124670.

Weyhenmeyer, Gesa A., and Jan Karlsson. 2009. “Nonlinear Response of Dissolved Organic Carbon Concentrations in Boreal Lakes to Increasing Temperatures.” *Limnology and Oceanography* 54 (6part2): 2513–19. https://doi.org/10.4319/lo.2009.54.6_part_2.2513.

Wilkinson, G. N., and C. E. Rogers. 1973. “Symbolic Description of Factorial Models for Analysis of Variance.” *Applied Statistics* 22 (3): 392. https://doi.org/10.2307/2346786.

Wing, Max Kuhn Contributions from Jed, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, et al. 2019. *Caret: Classification and Regression Training* (version 6.0-84). https://CRAN.R-project.org/package=caret.

Xu, Qifa, Kai Deng, Cuixia Jiang, Fang Sun, and Xue Huang. 2017. “Composite Quantile Regression Neural Network with Applications.” *Expert Systems with Applications* 76 (June): 129–39. https://doi.org/10.1016/j.eswa.2017.01.054.

Xu, Qifa, Xi Liu, Cuixia Jiang, and Keming Yu. 2016. “Quantile Autoregression Neural Network Model with Applications to Evaluating Value at Risk.” *Applied Soft Computing* 49 (December): 1–12. https://doi.org/10.1016/j.asoc.2016.08.003.