The following publication Y. Wang, I. W. -H. Ho, Y. Chen, Y. Wang and Y. Lin, "Real-Time Water Quality Monitoring and Estimation in AIoT for Freshwater Biodiversity Conservation," in IEEE Internet of Things Journal, vol. 9, no. 16, pp. 14366-14374, 15 Aug.15, 2022 is available at https://dx.doi.org/10.1109/JIOT.2021.3078166.

Real-time Water Quality Monitoring and Estimation in AIoT for Freshwater Biodiversity Conservation

Yuhao Wang, Ivan Wang-Hei Ho, *Senior Member*, *IEEE*, Yang Chen, Yuhong Wang, Yinghong Lin

***Abstract*—Deteriorating water quality leads to the freshwater biodiversity crisis. The interrelationships among water quality parameters and the relationships between these parameters and taxa groups are complicated in affecting biodiversity. Nevertheless, due to the limited types of Internet of Things (IoT) sensors available on the market, a large number of chemical and biological parameters still rely on laboratory tests. With the latest advancement in artificial intelligence and the IoT (AIoT), this technique can be applied to real-time monitoring of water quality, and further conserving biodiversity. In this paper, we conducted a comprehensive literature review on water quality parameters that impact the biodiversity of freshwater and identified the top-10 crucial water quality parameters. Among these parameters, the interrelationships between the IoT measurable parameters and IoT unmeasurable parameters are estimated using a general regression neural network model and a multivariate polynomial regression model based on historical water quality monitoring data. Conventional field water sampling and in-lab experiments, together with the developed IoT-based water quality monitoring system were jointly used to validate the estimation results along an urban river in Hong Kong. The general regression neural network model can successfully distinguish the abnormal increase of parameters against normal situations. For the multivariate polynomial regression model of degree eight, the coefficients of determination results are 0.89, 0.78, 0.87, and 0.81 for NO3-N, BOD$_5$, PO4, and NH3-N, respectively. The effectiveness and efficiency of the proposed systems and models were validated against laboratory results and the overall performance is acceptable with most of the prediction errors smaller than 0.2mg/L, which provides insights into how AIoT techniques can be applied to pollutant discharge monitoring and other water quality regulatory applications for freshwater biodiversity conservation.**

***Index Terms*— Artificial intelligence models, Freshwater biodiversity, Internet of Things, Top-10 crucial water quality parameters, Water quality monitoring, Water quality parameter estimation**

I. Introduction

Freshwater makes up only 0.01% of the global water but it supports at least 100,000 (almost 6%) of all recorded biological species [1]. Since aquatic species spend at least part of their lifetimes in water bodies, water quality directly affects their living condition, composition, distribution, and diversity [2, 3]. Relationships between water quality and biodiversity are well recorded in a variety of studies [4-7], and the deterioration of water quality is believed to be one of the factors contributing to the rapid decline in global aquatic biodiversity. To conserve freshwater biodiversity, water quality should be continuously monitored and evaluated. Conventional laboratory water quality tests, however, can only provide sparse data due to financial and time limits.

Over the last decade, the IoT has become a fresh and promising technique in water quality monitoring [9], especially for agriculture [8] and waste management [10]. The IoT technology is to connect traditional objects to the Internet to make things smart by utilizing technologies such as sensors, wireless communications and networking, cloud computing, and so on. According to Mohammadi [11], IoT-based services will contribute more than $2.7 trillion to global economics annually in 2025. Nevertheless, real-time water quality monitoring of diverse Physical-Chemical-Biological (PCB) parameters still remains a great challenge, primarily due to the limited types of sensors available on the market. Consequently, a large number of chemical and biological parameters still rely on laboratory tests, which are time-consuming and not cost-effective. To address these problems, this research aims to achieve the following objectives:

(1) Identify crucial water quality parameters that affect freshwater biodiversity;

(2) Identify those water quality parameters that can be measured with available IoT sensors and develop an IoT system to measure these parameters simultaneously;

(3) Develop artificial intelligence (AI) models to estimate parameters that cannot be measured by current IoT sensors using IoT-measurable parameters, based on a large historical water quality monitoring database.

(4) Evaluate the AI models using a case study.

The rest of the paper is organized as follows. In Section II, relationships between water quality parameters and biodiversity over the last two decades are reviewed and the top-10 crucial water quality parameters are identified. In Section III, a

This work was supported in part by the Research Impact Fund (Project No.R5007-18) established under the University Grant Committee (UGC) of the Hong Kong Special Administrative Region (HKSAR), China and in part by the research project (E-PolyU502/16) funded by the Research Grant Council (RGC) of the Hong Kong Special Administrative Region Government.

Y. Wang and I. W.-H. Ho are with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong (e-mail: yuhao1995.wang@connect.polyu.hk, ivanwh.ho@polyu.edu.hk)

Y. Chen, Y. Wang and Y. L are with the Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong (e-mail: yuhong.wang@polyu.edu.hk).

framework for unmeasurable water quality parameter estimation is proposed, a developed water quality monitoring system based on IoT technologies is introduced, and AI models to estimate unmeasurable water quality parameters are illustrated. In Section IV, the IoT system and estimation models are evaluated in a case study. Section V concludes the study and provides further recommendations.

II. Literature Review

A comprehensive literature review was conducted to identify water quality parameters that affect freshwater biodiversity. The following combinations of search terms were used in google scholar to find relevant studies between 2000 and 2020: (Fish OR benthic macroinvertebrates OR (Ephemeroptera, Plecoptera, and Trichoptera, EPT) OR freshwater/aquatic insects OR freshwater/aquatic macrophyte OR phytoplankton OR zooplankton) AND (biodiversity/diversity) AND (water quality parameters). A total of 90 papers were identified[2], which discuss important water quality parameters that affect freshwater biodiversity, organism distribution, and species composition.

According to [12], water quality parameters can be grouped into three categories: physical, chemical, and biological. Based on the search results, 33 physical-chemical-biological water quality parameters were identified as core factors, including 7 physical, 25 chemical, and 1 biological (E.coli). Apart from the three categories, another 5 hydrological parameters are also often discussed, including altitude, flow velocity, discharge, depth, and width. These parameters may partially contribute to the physical-chemical results.

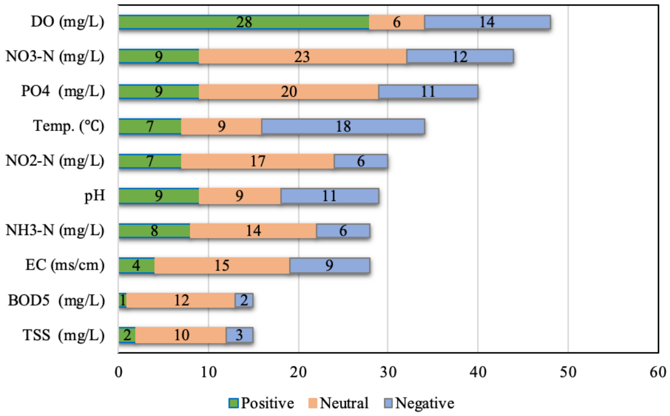


Fig. 1. The top-10 important water quality parameters affecting freshwater biodiversity (References can be referred to the supplementary material[2]).

*Neutral means the positive/negative relationship between the parameter and biodiversity is not clear, or it just shows the importance of the parameter affecting the species distribution and composition, or there are two completely different correlations for different families or species.

Based on occurrence frequencies, the top-10 critical parameters are listed in Fig. 1, consisting of 3 physical and 7 chemical parameters. The physical parameters include temperature, EC (electrical conductivity), and TSS (total suspended solids). Seven chemical parameters are frequently emphasized in a large body of literature, including DO (dissolved oxygen), NO3-N, PO4, NO2-N, pH, NH3-N, and BOD5 (5-day biological oxygen demand). Total nitrogen (TN) was classified into its main forms: NH3-N, NO3-N, NO2-N [13, 14]. Total phosphorus (TP) can be divided into dissolved phosphorus (DP) and particulate phosphorus (PP) [15]. However, since PP is seldom analyzed in the lab [15], the dissolved phosphorus PO4 is used as a surrogate [13]. Also shown in Fig. 1 are the roles of physicochemical properties in literature. Note that the roles vary from study to study because the responses of different species to these parameters are different.

The 90 reviewed papers were divided into two groups: 67 mainly concerning animals (fish, benthic macroinvertebrates, EPT, aquatic insect, and zooplankton) while 23 mainly concerning plants (aquatic macrophyte and phytoplankton). The proportion of papers that discuss the correlations between water quality parameters and animal/plant biodiversity are summarized in Fig. 2. As shown in Fig. 2, the same parameters may play different roles in animal and plant groups. For instance, DO generally plays a positive role in enhancing freshwater animal biodiversity, while the effect is not obvious for plants. The relationship between water quality parameters and freshwater biodiversity is complicated and sensitive. To achieve maximum diversity, multiple parameters should be kept in a suitable range. For example, rotifer requires optimal nutrient and temperature conditions, and a favorable DO range to achieve a higher diversity [16]. To balance the diversity among the animal and plant communities, some particular parameters should be closely monitored. For instance, nutrients are likely to promote plant diversity [14, 17] while curbing the growth of animals [14, 18]. Therefore, to provide the early warning of water quality conditions, and conserve freshwater biodiversity, real-time monitoring of water quality parameters is a necessity.

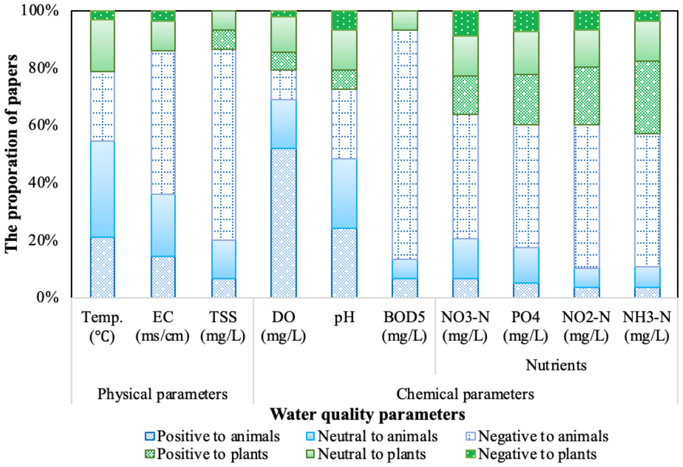


Fig. 2. The proportion of papers showing the correlation between water quality parameters and freshwater biodiversity.

Over the past decades, a large number of neural network models have been developed to estimate and predict water quality indicators. The effectiveness of this technique is exemplified by Gazzaz *et al.* [19], Han *et al.* [20], Zhang *et al.* [21], and a case study in India [22]. The corresponding models

[2] Reference list can be referred to the supplementary material [Online]. Available: https://github.com/wyhpolyu2020/WaterQualityinAIoT

include the feed-forward three-layer neural network, empirical neural network efficient self-organizing RBF neural network, and simple artificial neural network (ANN). In addition, a hybrid neural network was used for time series prediction of water quality in [23].

A variety of statistical models are also used in water quality parameter estimation including multiple linear regression [24]. Abyaneh [25] compared the performance between statistical models (e.g., multivariate linear regression model) and ANN models in estimating the biochemical oxygen demand (BOD) and chemical oxygen demand (COD), which shows ANN performs better. However, apparently, the linear model might not be able to fit the interrelationship well.

Most of the previous models aim for single parameter estimation and prediction. The difficulties and challenges of data collection and the feasibility of these models for IoT and other industrial applications have not been considered.

III. Research Methodology

The research methodology is introduced in this section. The framework for estimating unmeasurable water quality parameters is proposed in subsection A. A developed water quality monitoring system based on IoT technologies is presented in subsection B. The statistical feature analysis and two AI models to estimate unmeasurable water quality parameters are illustrated in subsection C, D, and E, respectively.

A. Problem Analysis

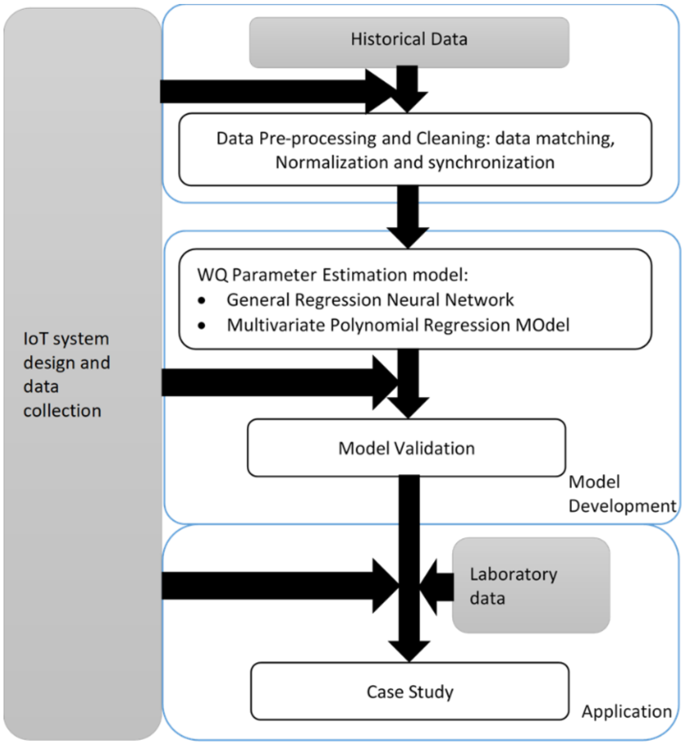

Historical Data
Data Pre-processing and Cleaning: data matching, Normalization and synchronization
WQ Parameter Estimation model:
• General Regression Neural Network
• Multivariate Polynomial Regression MOdel
IoT system design and data collection
Model Validation
Model Development
Laboratory data
Case Study
Application

Fig. 3. A three-step data-driven framework for water quality parameter estimation.

After the selection of the top-10 crucial water quality parameters in the previous section, a market survey was conducted to identify available sensors for developing the IoT-based water quality monitoring system. Five different types of sensors were found (i.e., total dissolved solids, pH, temperature, dissolved oxygen, and electrical conductivity). The other five parameters (i.e., NO3-N, PO4, NO2-N, NH3-N, and 5-day biological oxygen demand), thus as unmeasurable, will be estimated in the following process. Note that total dissolved solids (TDS) are not equivalent to TSS, but TDS is chosen due to the availability of the IoT sensor.

A three-step data-driven framework is proposed, as shown in Fig. 3. The three highlighted components are data inputs from three different sources. Firstly, historical data are cleaned and pre-processed. The models for estimating unmeasurable parameters are developed in the next step. The last step is the implementation of the case study and evaluation of the proposed models adopting IoT sensor data and laboratory data.

B. Development of the IoT Water Quality Monitoring System

Five different types of sensors were identified on the market. In addition to these sensors, the Wemos D1 Mini chip and a multiplexer were used in building this IoT system. The Wemos D1 Mini Chip is a portable integrated chip that has ESP8266 Wi-Fi and Arduino functions. The casual role of the multiplexer in the IoT system is that most of the sensors only enable analog output while only one analog input pin is available in this small chip so the multiplexer is used for pin expansion. The IoT server is based on the ThingsBoard professional edition [26].

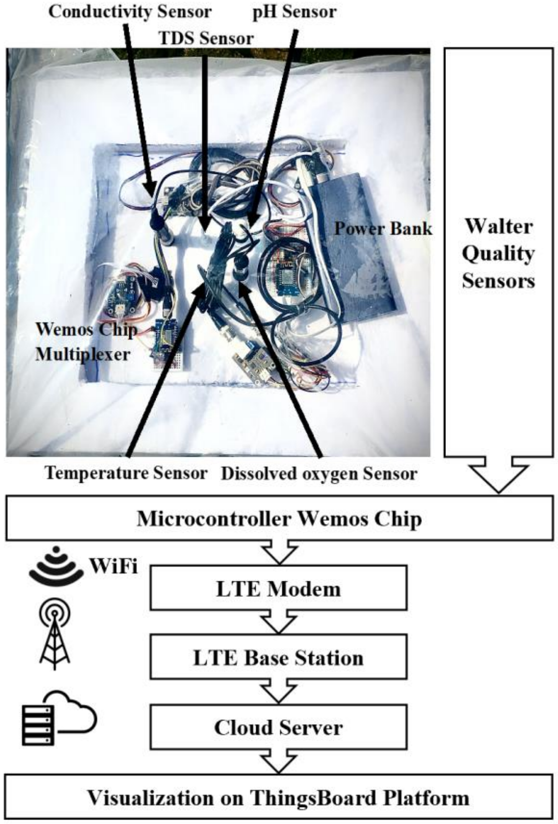


Fig. 4. The developed water quality monitoring IoT system based on a Wemos chip and a multiplexer.

The system is embedded in a foam board for floating on the water surface and covered with a plastic film for waterproofing.

After the sketch of the system was made and the connections were drawn, a program in the chip was developed so that it could collect and send the five-parameter data to the web-server continuously and simultaneously via Wi-Fi to 5G gateway and upload to the Internet. The whole system and the server dashboard are shown in Fig. 4 and Fig. 5 respectively.

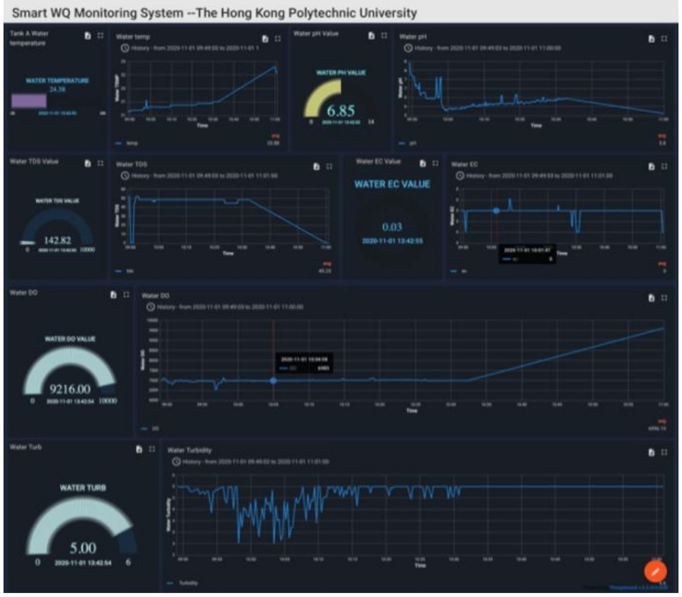


Fig. 5. The ThingsBoard web-server dashboard of the developed IoT system.

C. Statistical Feature Analysis

Prior to data modeling, the statistical features of the historical data were examined through descriptive analysis. Except for the essential correlation analysis, count, mean, standard deviation, minimum value, maximum value, and each quartile value of the raw data set are analyzed in this step. These data potentially reveal the relationships among 10 crucial parameters, and their central tendencies as well as the dispersion. The Pearson's correlation coefficient in (1) was applied to investigate the potential relationships among these parameters. The Pearson's correlation can measure the magnitude of a linear relationship between a paired data set [27].

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}} \,. \qquad (1)$$

The correlation coefficient *r* ranges from -1 to 1. In general, a larger absolute value of *r* indicates a stronger linear relationship. The positive value implies a positive linear correlation between the compared data set and vice versa. If *r* = 0, it means there is no linear correlation.

D. General Regression Neural Network

Artificial Neural Network (ANN) models were utilized in estimating unmeasurable water quality parameters. General regression neural network (GRNN) is a single-pass associative memory feed-forward type ANN and was employed in this study due to its quick training approach and high accuracy. However, the disadvantage of GRNN is its growth of the hidden layer size [30]. MSE (mean squared error) generally is a conspicuous measurement of GRNN. According to an existing study [30], GRNN has less training time and higher accuracy than back-propagation ANN.

E. Multivariate Polynomial Regression Model

According to Ostertagová [28], polynomial regression (PR) is a special case of multiple regression in the machine learning domain. It fits the data using least-square methods, which could minimize the variance of the unbiased coefficient estimators, under the Gauss-Markov theorem. The PR method is used when the response variable is non-linear and the general equation for PR is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_k x^k + \varepsilon, \qquad (2)$$

where ε denotes the random error term which follows normal distribution $\varepsilon \sim N(0, \sigma^2)$ andβ_i, x, y are coefficient parameter, independent variable, and dependent variable respectively.

When PR is applied to multiple regression variables, it could be regarded as multivariate polynomial regression (MPR) [29]. For example, the expression of a second-order MPR is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon, \; (3)$$

where β_i, β_{ii} and β_{ij} are called linear effect, quadratic effect, and interaction effect parameters separately. With an increase in the number of independent variables, the number of polynomial parameters will jump exponentially. MPR can also be represented in a matrix form which is employed in our model:

$$\boldsymbol{Y} = \boldsymbol{\beta X} + \boldsymbol{\varepsilon}. \qquad (4)$$

In this paper, the PolynomialFeatures transformer of the scikit-learn machine learning library in Python was used to construct the polynomial variables as input to linear regression to train on the non-linear functions. The major problem of MPR is multicollinearity because the parameters are highly likely to be interdependent on each other and this could cause poor performance on model fitting. Possible solutions to this problem are curvilinear distance analysis, Sammon's mapping, and kernel principal component analysis [29].

IV. Case Study and Results

A. Background

Lam Tsuen River is a major river in the Eastern New Territories in Hong Kong. From the Hong Kong government online database (https://data.gov.hk/en/), we utilized 30 years of historical river water quality data at different monitoring stations from 1986 to 2018 to develop the MPR and GRNN models for the estimation of the other five unmeasurable crucial parameters (i.e., NO3-N, PO4, NO2-N, NH3-N, and 5-day biological oxygen demand). A total of 34,650 pieces of historical data were obtained after data pre-processing for subsequent analysis. All of these 10 parameters are included in the historical data set. As mentioned before, because there is no TSS sensor available on the market, we used a TDS sensor instead. When developing the GRNN model and MPR model based on the historical data set, we obtained TDS data using total solids minus TSS.

B. *Deployment of the IoT system*

Four sampling sites along the Lam Tsuen River were selected as experiment sites, as shown in Fig. 6. The developed IoT water quality monitoring system was set up on these four sampling sites to collect the data of the five measurable parameters (i.e., TDS, pH, temperature, dissolved oxygen, and electrical conductivity). Moreover, water samples were collected and analyzed using conventional lab methods to validate the models. The first sampling site is located at the most upstream natural place. The second sampling site is located almost 500 m downstream of the first site, which is just flowing through a small village and farmland. Additionally, the third and fourth sites are the upstream and downstream of an ecological restoration site of the lower reach. The lower reach, called Lower Lam Tsuen River, passes through the urban area of Tai Po [31]. Restoration efforts began in 2016, and the ecologically ‘enhanced’ section (N22°27'0.492" E114°9'30.478") spans approximately 40 meters. Concerning the top-10 important water quality parameters, the measurement methods can be referred to [32]. Due to the nitrification process, the volume of NO2-N was not measured in this research [33].

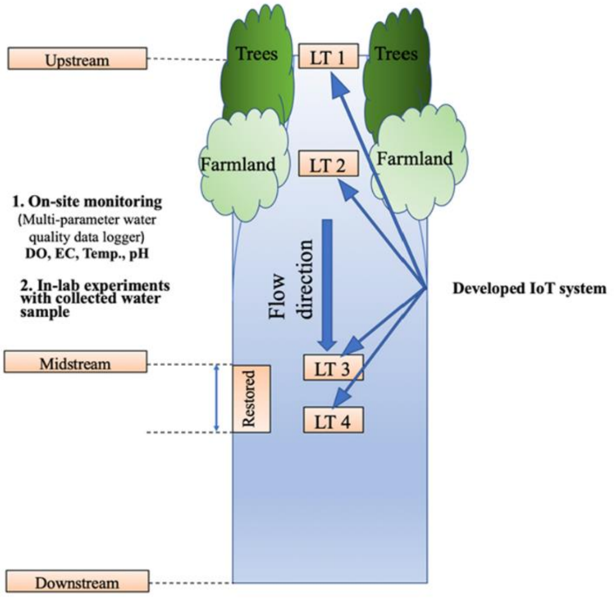


Fig. 6. Sketch of the four sampling sites along the Lam Tsuen River.

The sampling frequency of the developed IoT system is around 5 seconds so that about 360 groups of data are collected at each sampling site. Because of only one group of data available in the laboratory result, we use the mean value of the IoT parameter data as the input of validation. Finally, the water quality results listed in Table I are used to evaluate the model performance. The five measurable parameters are the mean value of the data obtained by the IoT system at each location whereas the four unmeasurable parameters are laboratory results of the collected water sample.

TABLE I
WATER QUALITY PARAMETER RESULTS FOR MODEL VALIDATION

Sampling sites	LT 1	LT 2	LT 3	LT 4
DO (mg/L)	8.46	8.93	8.99	8.12
NO_3-N (mg/L)	0	0.4	1.4	1.7
PO_4 (mg/L)	0.03	0.1	0.22	0.26
Temp. (℃)	20.6	21.6	25.4	25.7
pH	7.2	7.38	6.31	6.77
NH_3-N (mg/L)	0.01	0.01	0.32	0.19
EC (ms/cm)	0.037	0.042	0.136	0.141
BOD_5 (mg/L)	0.22	0.47	1.55	1.37
TDS (mg/L)	48.08	18.7	136.02	142.82

C. *Statistical Feature Analysis of Historical Data*

Some basic statistic description numbers of the 10 crucial parameters from the historical dataset are shown in Table II. Fig. 7 demonstrates the overall Pearson’s correlation matrix. The larger the circle is, the stronger the correlation there is. The motivation behind using the Pearson’s correlation analysis is that before utilizing the following models, we need to check the strength of linear correlation between the independent variables to avoid perfect linear correlated variables. The water dissolved oxygen is almost negatively correlated with all the other parameters while positively with pH, and the strengths of correlations are comparatively high. The strengths of dissolved oxygen and orthophosphate phosphorus are relatively greater than others. The strengths among 5-day biochemical oxygen demand, ammonia-nitrogen, and orthophosphate phosphorus are very large, and they are positively correlated with each other while negatively correlated with dissolved oxygen.

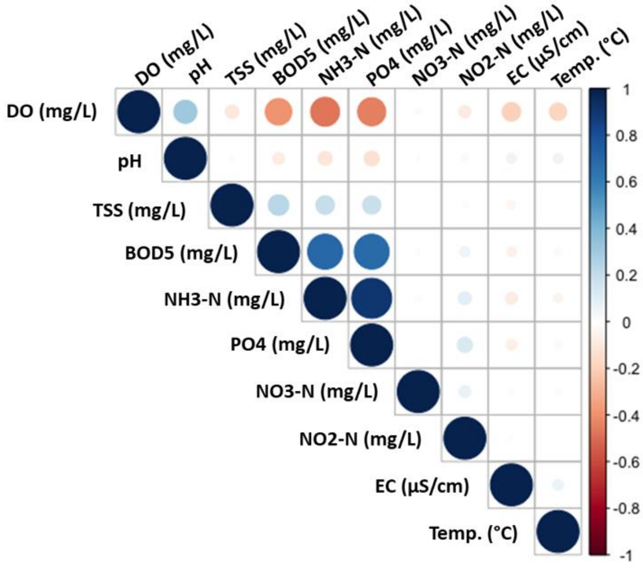


Fig. 7. The correlation analysis of the top-10 crucial parameters.

TABLE II
BASIC STATISTIC DESCRIPTION NUMBERS OF THE TOP-10 CRUCIAL PARAMETERS FROM THE HISTORICAL DATA SET

	EC (μS/cm)	DO (mg/L)	pH	NO3-N (mg/L)	NO2-N (mg/L)	Temp. (°C)	BOD5 (mg/L)	PO4 (mg/L)	TSS (mg/L)	NH3-N (mg/L)
count	34650.00	34650.00	34650.00	34650.00	34650.00	34650.00	34650.00	34650.00	34650.00	34650.00
mean	2334.68	8.10	7.43	1.12	0.08	23.44	2.84	0.17	18.49	0.47
std	7345.45	1.78	0.46	8.27	0.25	4.52	4.74	0.36	85.95	1.19
min	2.00	0.30	4.60	0.00	0.00	9.30	0.00	0.00	0.00	0.00
25%	66.00	7.50	7.20	0.33	0.00	19.90	0.50	0.03	1.50	0.03
50%	100.00	8.30	7.40	0.67	0.01	23.90	1.20	0.06	3.20	0.07
75%	188.00	9.10	7.60	1.10	0.05	27.00	3.30	0.14	8.90	0.38
max	47824.00	20.80	11.60	480.00	7.91	36.20	92.10	4.30	2300.00	16.00

D. General Regression Neural Network

As mentioned in the previous section, NO2-N cannot be validated using laboratory data and thus the other four unmeasurable parameters (i.e., NO3-N, PO4, NH3-N, and 5-day biological oxygen demand) are estimated in the GRNN model. Before the modeling process, the input data were standardized using the scale function of the scikit-learn machine learning library (centered to the mean and component-wise scaled to unit variance).

The GRNN model is a network with five standardized input variables, a hidden layer with 30 processing neurons. The loss function is MSE and the optimizer is Adam. The historical data set is separated into a 90% training set and 10% validation set. The modeling work was accomplished on TensorFlow version 2.3.0. Fig. 8 shows the goodness of fit of the standardized NO3-N. We can see from the figure that the model approximately fits with the dataset except for some local peaks. The total MSEs of this model are 0.9148, 0.9331, 1.1060, and 1.0125 for BOD_5, NH3-N, NO3-N, and PO4, respectively. Since this is a regression problem rather than a classification problem, the MSE values are relatively large. Another reason is that the Nitrate-Nitrogen has a small mean value of 1.12 mg/L but an extremely large maximum value of 480 mg/L according to Table II. The 75% quartile value is 1.10 mg/L which is less than the mean value. This indicates that the distribution of Nitrate-Nitrogen data has a long tail skewness along the x-axis direction, thus generating larger errors.

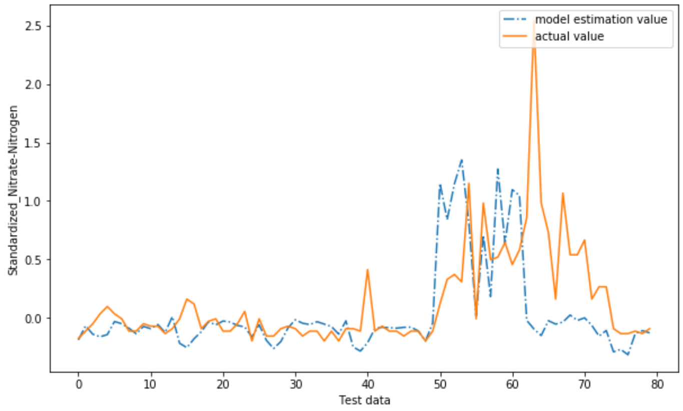


Fig. 8. The fitting performance of the GRNN model.

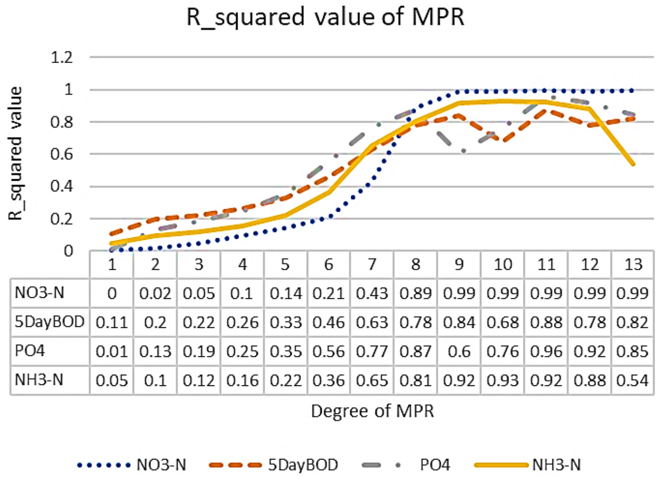
However, Fig. 8 indicates that when the actual data remain steady, the model estimation results also remain relatively constant with good performance. On the other hand, when there is a sharp increase in the value of the actual data, the model estimation result has a jump as well. This suggests that our model has strong potential in classifying the normal situation against the abnormal cases. Such a classifier can be applied to pollutant discharge monitoring and other water quality regulatory applications for conserving biodiversity.

E. Multivariate Polynomial Regression Model

Based on the correlation analysis, among the IoT measurable parameters, the largest correlation value is 0.3 which is between DO and pH. This means the absence of perfect multicollinearity (i.e., an exact but no stochastic linear relationship) between each independent variables. However, such moderate multicollinearity does not affect the precision of the predictions and the goodness of fit statistics, and hence the problem can be ignored in the MPR model. The MPR models are fitted using the historical data set. The coefficient of determination is often denoted r^2 which is applied to evaluate the performance of the MPR model.

$$r^2 = 1 - \frac{SS_{res}}{SS_{tot}}, \qquad (5)$$

where SS_{res} is the sum of squares of residuals and SS_{tot} is the total sum of squares according to each value and mean value. We use r^2 to examine the fitting performance between each unmeasurable parameter and the 5 measurable parameters.

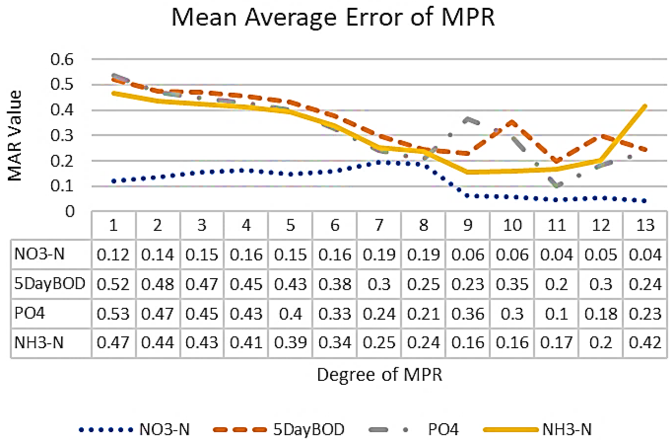


	1	2	3	4	5	6	7	8	9	10	11	12	13
NO3-N	0	0.02	0.05	0.1	0.14	0.21	0.43	0.89	0.99	0.99	0.99	0.99	0.99
5DayBOD	0.11	0.2	0.22	0.26	0.33	0.46	0.63	0.78	0.84	0.68	0.88	0.78	0.82
PO4	0.01	0.13	0.19	0.25	0.35	0.56	0.77	0.87	0.6	0.76	0.96	0.92	0.85
NH3-N	0.05	0.1	0.12	0.16	0.22	0.36	0.65	0.81	0.92	0.93	0.92	0.88	0.54

Fig. 9. R_squared value of multivariate polynomial regression.

Fig. 9 shows that there is a gradual increase in the value of the coefficient of determination r^2 with the increasing of the

MPR degree. It also indicates that the growth trend has three stages for those four unmeasurable parameters: steady increase before the first five degrees, sharp rising from degree 6 to 9, and steady again after degree 10 which could be due to overfitting. The fluctuations of PO4 and 5DayBOD after degree 10 could be due to the fitting ability changes with the increase of the degree. Each dependent variable has a different sensitivity to the MPR degree because of its statistical characteristics. Overall, the model still has a growing fitting ability trend with a rising MPR degree. The observed sharp rising in r^2 could be attributed to the increasing fitting ability of the MPR model.



	1	2	3	4	5	6	7	8	9	10	11	12	13
NO3-N	0.12	0.14	0.15	0.16	0.15	0.16	0.19	0.19	0.06	0.06	0.04	0.05	0.04
5DayBOD	0.52	0.48	0.47	0.45	0.43	0.38	0.3	0.25	0.23	0.35	0.2	0.3	0.24
PO4	0.53	0.47	0.45	0.43	0.4	0.33	0.24	0.21	0.36	0.3	0.1	0.18	0.23
NH3-N	0.47	0.44	0.43	0.41	0.39	0.34	0.25	0.24	0.16	0.16	0.17	0.2	0.42

Fig. 10. Mean average errors of multivariate polynomial regression.

Fig. 10 reveals that with the increase in degree, the Mean Absolute Errors (MAEs) decrease steadily with some fluctuations after degree 8 for the three parameters other than NO3-N. NO3-N has the smallest MAE when the model degree is 1. A possible explanation for this abnormal result might be that NO3-N has a relatively large standard deviation according to Table II and a significant maximum value compared with its quartile values. This might lead to a flat distribution which means a smaller MAE when the degree of the MPR model is small.

Table III compares the estimated results with the lab results at four different sampling sites. Most of the errors are smaller than 0.2 mg/L which suggests that the satisfactory performance of the model. However, significant differences have been found in the BOD5 estimation. The error is significantly larger than other parameters while others are less than 0.2 mg/L except for NO3-N at the first site. This discrepancy may be attributed to the time interval between the sample was collected and tested. It is noticed that the errors in LT3 and LT4 are relatively smaller than LT1 and LT2. There are more pollutants downstream of the river, making the IoT sensors more sensitive. In addition, most data in the historical database used for model development were collected in more polluted sites than LT1 and LT2, where water is much cleaner. Therefore, the developed models and methods appear to be more suitable to monitor polluted river sites.

TABLE III
Comparison Between Model Results and Laboratory Results

	NO3-N (mg/L)			BOD5 (mg/L)		
Location	Model result	Lab result	error	Model result	Lab result	error
LT1	0.493	0.000	-0.493	2.141	0.220	-1.921
LT2	0.233	0.400	0.167	1.658	0.470	-1.188
LT3	1.312	1.400	0.088	1.306	1.550	0.244
LT4	1.725	1.700	-0.025	1.877	1.370	-0.507
	PO4 (mg/L)			NH3-N (mg/L)		
Location	Model result	Lab result	error	Model result	Lab result	error
LT1	0.150	0.030	-0.120	0.063	0.010	-0.053
LT2	0.170	0.100	-0.070	0.063	0.010	-0.053
LT3	0.184	0.220	0.036	0.371	0.320	-0.051
LT4	0.208	0.260	0.052	0.299	0.190	-0.109

V. Conclusion

Real-time water quality monitoring using IoT-connected sensors provides a promising method to monitor water quality for conserving freshwater biodiversity. However, monitoring a full spectrum of physical-chemical-biological parameters remains a challenge, due to the limited types of sensors available on the market.

This paper firstly identified crucial water quality parameters from the perspective of biodiversity conservation. A data-driven framework for estimating unmeasurable water quality parameters was subsequently proposed, and a real-time water quality monitoring IoT system was developed. Specifically, we have proposed the GRNN model and the MPR model to detect abnormal discharge of pollutants and estimate unmeasurable critical water quality parameters from measurable ones by the IoT system respectively. The GRNN model was found to be able to distinguish abnormal increase of parameters against normal situations, while the MPR model of degree eight has coefficient of determination values of 0.89, 0.78, 0.87, and 0.81 for NO3-N, BOD_5, PO4, and NH3-N, respectively. We have also evaluated the proposed systems and models via an experimental study along the Lam Tsuen River in Hong Kong. The performance of the proposed models was validated against laboratory results, and the overall performance appears to be acceptable and adequate, with most of the error values less than 0.2 mg/L for NO3-N, PO4, and NH3-N estimation. In addition, the prediction performs better at river sites with a relatively higher level of pollutants.

In the long term, the proposed framework needs to be further enhanced and fine-tuned to gain a good balance between accuracy and timeliness. Further machine learning techniques will also be studied for real-time water quality parameter estimation. For example, recurrent neural networks (e.g., long short-term memory) can be used for sequential water quality data prediction.

References

[1] D. Dudgeon *et al.*, "Freshwater biodiversity: importance, threats, status and conservation challenges," *Biological reviews,* vol. 81, no. 2, pp. 163-182, 2006.

[2] F. O. Arimoro and R. B. Ikomi, "Ecological integrity of upper Warri River, Niger Delta using aquatic insects as bioindicators," *Ecological indicators,* vol. 9, no. 3, pp. 455-461, 2009.

[3] I. N. Khan, "Assessment of water pollution using diatom community structure and species distribution—a case study in a tropical river basin," *Internationale Revue der gesamten Hydrobiologie und Hydrographie,* vol. 75, no. 3, pp. 317-338, 1990.

[4] M. Azrina, C. Yap, A. R. Ismail, A. Ismail, and S. Tan, "Anthropogenic impacts on the distribution and biodiversity of benthic macroinvertebrates and water quality of the Langat River, Peninsular Malaysia," *Ecotoxicology and environmental safety,* vol. 64, no. 3, pp. 337-347, 2006.

[5] K. M. Debjit, K. Anilava, and S. Subrata, "Water quality parameters and fish biodiversity indices as measures of ecological degradation: a case study in two floodplain lakes of India," *Journal of Water Resource and Protection,* vol. 2010, 2010.

[6] S. Palleyi and C. Panda, "Influence of water quality on the biodiversity of phytoplankton in Dhamra river Estuary of Odisha Coast, Bay of Bengal," *Journal of Applied Sciences and Environmental Management,* vol. 15, no. 1, 2011.

[7] T. Prommi and A. Payakka, "Aquatic insect biodiversity and water quality parameters of streams in Northern Thailand," *Sains Malaysiana,* vol. 44, no. 5, pp. 707-717, 2015.

[8] A. A. Pranata, J. M. Lee, and D. S. Kim, "Towards an IoT-based water quality monitoring system with brokerless pub/sub architecture," in *2017 IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN)*, 2017: IEEE, pp. 1-6.

[9] M. S. U. Chowdury *et al.*, "IoT based real-time river water quality monitoring system," *Procedia Computer Science,* vol. 155, pp. 161-168, 2019.

[10] M. V. Ramesh *et al.*, "Water quality monitoring and waste management using IoT," in *2017 IEEE Global Humanitarian Technology Conference (GHTC)*, 2017: IEEE, pp. 1-7.

[11] M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani, "Deep learning for IoT big data and streaming analytics: A survey," *IEEE Communications Surveys & Tutorials,* vol. 20, no. 4, pp. 2923-2960, 2018.

[12] N. H. Omer, "Water Quality Parameters," in *Water Quality-Science, Assessments and Policy*: IntechOpen, 2019.

[13] J. Aazami, A. Esmaili-Sari, A. Abdoli, H. Sohrabi, and P. J. Van den Brink, "Monitoring and assessment of water health quality in the Tajan River, Iran using physicochemical, fish and macroinvertebrates indices," *Journal of Environmental Health Science and Engineering,* vol. 13, no. 1, p. 29, 2015.

[14] K. Luo *et al.*, "Impacts of rapid urbanization on the water quality and macroinvertebrate communities of streams: A case study in Liangjiang New Area, China," *Science of The Total Environment,* vol. 621, pp. 1601-1614, 2018.

[15] B. M. Weigel and D. M. Robertson, "Identifying biotic integrity and water chemistry relations in nonwadeable rivers of Wisconsin: toward the development of nutrient criteria," *Environmental Management,* vol. 40, no. 4, pp. 691-708, 2007.

[16] B. Padmanabha and S. Belagali, "Comparative study on population dynamics of rotifers and water quality index in the lakes of Mysore," *Nature, Environment and Pollution Technology*, vol. 5, no. 1, pp. 107-109, 2006.

[17] C.-C. Sun et al., "Seasonal variation of water quality and phytoplankton response patterns in Daya Bay, China," *International Journal of Environmental Research and Public Health*, vol. 8, no. 7, pp. 2951-2966, 2011.

[18] C.-B. Hsu et al., "Biodiversity of constructed wetlands for wastewater treatment," *Ecological Engineering*, vol. 37, no. 10, pp. 1533-1545, 2011.

[19] N. M. Gazzaz, M. K. Yusoff, A. Z. Aris, H. Juahir, and M. F. Ramli, "Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors," *Marine pollution bulletin*, vol. 64, no. 11, pp. 2409-2420, 2012.

[20] H.-G. Han, Q.-l. Chen, and J.-F. Qiao, "An efficient self-organizing RBF neural network for water quality prediction," *Neural Networks*, vol. 24, no. 7, pp. 717-725, 2011.

[21] Y. Zhang, J. Pulliainen, S. Koponen, and M. Hallikainen, "Application of an empirical neural network to surface water quality estimation in the Gulf of Finland using combined optical data and microwave data," *Remote sensing of environment*, vol. 81, no. 2-3, pp. 327-336, 2002.

[22] K. P. Singh, A. Basant, A. Malik, and G. Jain, "Artificial neural network modeling of the river water quality—a case study," *Ecological Modelling*, vol. 220, no. 6, pp. 888-895, 2009.

[23] D. Ö. Faruk, "A hybrid neural network and ARIMA model for water quality time series prediction," *Engineering applications of artificial intelligence*, vol. 23, no. 4, pp. 586-594, 2010.

[24] K. S. D. Krishnan and P. Bhuvaneswari, "Multiple linear regression based water quality parameter modeling to detect hexavalent chromium in drinking water," in 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2017: IEEE, pp. 2434-2439.

[25] H. Z. Abyaneh, "Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters," *Journal of Environmental Health Science and Engineering,* vol. 12, no. 1, p. 40, 2014.

[26] T. ThingsBoard. https://thingsboard.io (accessed 2020).

[27] D. Wu, H. Mohammed, H. Wang, and R. Seidu, "Smart Data Analysis for Water Quality in Catchment Area Monitoring," in 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 2018: IEEE, pp. 900-908.

[28] E. Ostertagová, "Modelling using polynomial regression," *Procedia Engineering*, vol. 48, pp. 500-506, 2012.

[29] P. Sinha, "Multivariate polynomial regression in data mining: methodology, problems and solutions," *International Journal of Scientific and Engineering Research,* vol. 4, no. 12, pp. 962-965, 2013.

[30] A. J. Al-Mahasneh, S. G. Anavatti, and M. A. Garratt, "Review of applications of generalized regression neural networks in identification and control of dynamic systems," arXiv preprint arXiv:1805.11236, 2018.

[31] (2017). River Water Quality in Hong Kong in 2016. [Online] Available: http://wqrc.epd.gov.hk/pdf/water-quality/annual-report/RiverReport2016eng.pdf

[32] Y. W. Yang Chen, Bernadette Chia, Dawei Wang, "An Upstream-downstream Water Quality Comparison of Restored Urban Drainage Channels," 2020.

[33] B. Kholdebarin and J. Oertli, "Effect of pH and ammonia on the rate of nitrification of surface water," *Journal (Water Pollution Control Federation)*, pp. 1688-1692, 1977.

Yuhao Wang was born in Shuangyashan, Heilongjiang, China in 1995. He received the B.Eng. degree in traffic engineering from Southeast University, China, in 2018 and the M.S. degree in transport from Imperial College London in 2019. He is currently working toward the Ph.D. degree with The Hong Kong Polytechnic University, Kowloon, Hong Kong, focusing on intelligent transport system and smart pavement. His work on a smart, integrated road pavement and drainage system for stormwater storage, de-icing, dust suppression, and cooling received the Gold Medal with the Organizer's Choice Award in the International Invention Innovation Competition in Canada (iCAN) in 2020.

Ivan Wang-Hei Ho (M’10–SM’18) received the B.Eng. and M.Phil. degrees in information engineering from The Chinese University of Hong Kong, Hong Kong, in 2004 and 2006, respectively, and the Ph.D. degree in electrical and electronic engineering from the Imperial College London, London, U.K., in 2010. He was a Research Intern with the IBM Thomas J. Watson Research Center, Hawthorne, NY, USA, and a Postdoctoral Research Associate with the System Engineering Initiative, Imperial College London. In 2010, he cofounded P2 Mobile Technologies Ltd., where he was the Chief Research and Development Engineer. He is currently an Assistant Professor with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong

Kong. His research interests include wireless communications and networking, specifically in vehicular networks, intelligent transportation systems (ITS), and Internet of things (IoT). He primarily invented the MeshRanger series wireless mesh embedded system, which received the Silver Award in Best Ubiquitous Networking at the Hong Kong ICT Awards 2012. His work on indoor positioning and IoT also received the Gold Medal at the International Trade Fair Ideas and Inventions New Products (iENA) in Germany in 2019, and the Gold Medal with the Organizer's Choice Award in the International Invention Innovation Competition in Canada (iCAN) in 2020. He is currently an Associate Editor for the IEEE Access and IEEE Transactions on Circuit and Systems II, and was the TPC Co-Chair for the PERSIST-IoT Workshop in conjunction with ACM MobiHoc 2019 and IEEE INFOCOM 2020.

Yang Chen received the B.Eng. degree in civil engineering from Southwest Jiaotong University, China, in 2013 and the M.Sc. degree in technology management from The Hong Kong Polytechnic University, in 2018. She is currently pursuing the Ph.D. degree with the Hong Kong Polytechnic University, Kowloon, Hong Kong. She participates in the UNaLab project, focusing on urban biodiversity evaluation and conservation.

Yuhong Wang received his M.Eng. and B.Eng. degree at Tongji University, Shanghai, China in 1996, and the Ph.D. and MSc in civil engineering at the University of Kentucky, the USA in 2003 and 2001, respectively. He is currently a Professor with the Department of Civil and Environmental Engineering, Hong Kong Polytechnic University, Hong Kong. His research interest focuses on the new generation of urban infrastructure, which includes how to make future cities cleaner and more environmentally friendly, more resistant to floods, better serve urban residents, smarter, and how to promote biodiversity in the urban environment.

Yinghong Lin was born in Chongqing, China in 1997. She received the B.Eng. degree in water supply and drainage engineering from The Shandong Jianzhu University, China, in 2019 and the M.Sc. degree in environmental management and engineering from The Hong Kong Polytechnic University in 2020. She is currently working toward the Ph.D. degree with The Hong Kong Polytechnic University, Kowloon, Hong Kong, focusing on the urban stormwater management system through the pavement and sponge city.