# Correcting Systematic Underprediction of Biochemical Oxygen Demand in Support Vector Regression

Marvin X. C. Seow[1] and Alan D. Ziegler[2]

**Abstract:** Biochemical oxygen demand (BOD) is a variable that is missing or inaccurate in many water quality data sets because of difficulties in diluting highly polluted water samples. Machine learning algorithms, particularly support vector regression (SVR), are useful to build regression models to fill gaps in these data sets. The SVR can underpredict extreme-high values when they are few in number and underrepresented. This paper evaluates two methods, bootstrapping and data expansion, to mitigate the problem by increasing the proportion of extreme-high BOD in the data set before training the gap-filling model. Both methods were tested on the water quality data of Yuen Long Creek, Hong Kong, for the years 2000–2014. Both methods were effective in mitigating systematic underprediction and reducing their residual errors when the proportion of extreme-high values in the data set were increased from 3 to 30–40%. Both methods were useful for gap filling on BOD time series because extreme-high values are often the ones missing or inaccurate when highly polluted samples are diluted. **DOI: 10.1061/(ASCE)EE.1943-7870.0001243.** 



## Introduction

Biochemical oxygen demand (BOD), a widely used criterion for water quality assessment, is the measure of the quantity of dissolved oxygen (DO) used by microorganisms to decompose organic matter in water (Sawyer et al. 2002). It is fundamentally an indicator of organic pollution. Operationally, BOD is determined as the DO concentration absorbed by a sample maintained at temperature of 20°C for a fixed period (normally five days), expressed in $mg\,L^{-1}$, before nitrogen matter begins decomposing (Nagel et al. 1992; Sawyer et al. 2002). Occasionally, grossly polluted water samples with low DO concentration (<6 $mg\,L^{-1}$) must be diluted with specific amounts of water and nutrients before testing (Chiang et al. 2004). The addition of water maintains an adequate oxygen supply for complete decomposition to occur. The addition of nutrients maintains the bacterial decomposition rate. Because it is impossible to know the actual BOD beforehand, multiple samples of different dilution amounts have to be prepared according to a predicted BOD determined by the laboratory analyst (Sawyer et al. 2002).

Measurements of BOD are deemed inaccurate if the final DO concentration after the test falls below 1 $mg\,L^{-1}$ and is at least 2 $mg\,L^{-1}$ lower than the initial DO concentration (Rice et al. 2012). Inaccurate measurements frequently occur during the dilution process for several reasons, such as incorrect laboratory technique, unacceptable dilution water quality, and toxicity (Chiang et al. 2004). For example, ~10% of the BODs in the 2000–2014 water quality data set for Yuen Long Creek (Hong Kong), which are marked by blanks and inequality signs, and those in the data set studied in this paper are unreliable in part because of inaccurate dilution. In general, incomplete data sets (data gaps) pose problems for time-based monitoring studies aimed at improving river water quality management or identifying point sources. Therefore, new or improved techniques to fill gaps or "correct" inaccurate values are potentially quite valuable.

The BOD is influenced by several physical (e.g., pH, temperature, flow rate) and chemical/biological (e.g., anions, bacteria, metals) water quality variables (Džeroski et al. 2000; Singh et al. 2009; Udeigwe and Wang 2010). The (often) nonlinearity of the relationships among these parameters necessitates the use of sophisticated methods to fill gaps and to correct values. Machine learning approaches involving support vector regression (SVR), for example, may have superior predictive capability over simpler, traditional regression methods because they can model nonlinear relationships between variables. These relationships cannot be captured by linear regression methods (Singh et al. 2009; Lima et al. 2015). SVR adopts a kernel-based approach that can simultaneously reduce model dimensions and minimize prediction errors. It has been used to produce robust regression models for BOD prediction (e.g., Noori et al. 2012, 2015).

In some instances, the performance of SVR may be reduced by the systematic underprediction of extreme-high dependent values in a data set. This paper defines the problem of systematic underprediction as extreme-high values being consistently underestimated with a large negative bias, but the mean residual error of nonextreme values is near zero and they are not affected by any systematic artifact. In the case of the 2000–2014 data set for Yuen Long Creek, extreme-high BODs are subject to underprediction. Inherently, the systematic underprediction results from an SVR attempting to minimize both prediction errors and model complexity (Balfer and Bajorath 2015). As extreme-high values often represent only a small proportion of the data set, the SVR will algorithmically tolerate prediction errors in an effort to derive a sufficiently complex model that provides accurate predictions for the majority of the data—that is, the nonextreme values. The problem of systematic underprediction can be relevant to other machine learning and traditional regression techniques, but this paper limits the discussion to the SVR.

In the context of BOD prediction, the accurate prediction of extreme-high values is important because (1) they are associated with high levels of pollution and (2) they are the values most likely to

[1]Graduate Student, Graduate School of Science, Univ. of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan (corresponding author). ORCID: http://orcid.org/0000-0003-4930-9829. E-mail: xcmarvin@eps.s.u-tokyo.ac.jp

[2]Professor, Dept. of Geography, National Univ. of Singapore, 1 Arts Link, Kent Ridge, Singapore 117570. E-mail: adz@nus.edu.sg

Note. This manuscript was submitted on December 7, 2016; approved on February 7, 2017; published online on May 8, 2017. Discussion period open until October 8, 2017; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Environmental Engineering***, © ASCE, ISSN 0733-9372.

incur errors during dilution. Such systematic underprediction of extreme-high values has been reported in some SVR modeling literature, although ways to correct the underprediction issue have yet to be suggested. For instance, Garsole and Rajurkar (2015) and Granata et al. (2016) identified that the SVR tends to underpredict peak hydrological discharges. A pharmaceutical modeling study by Balfer and Bajorath (2015) noted that the SVR underpredicted the potency value of highly potent compounds. One approach to resolve systematic underprediction is to increase the proportion of extreme-high values in the data set prior to building the model. This can be done by replicating extreme-high values already in the data set. This approach has not been evaluated in any literature on SVR or other regression techniques. This paper evaluates the effectiveness of two novel replication methods based on bootstrapping and data expansion to the SVR modeling of the 2000–2014 BOD times series for Yuen Long Creek, Hong Kong's most polluted river (Environmental Protection Department of Hong Kong 2014).

## Material and Methods

### Study River

The Yuen Long Creek is situated in subtropical Hong Kong. It runs north from the central hills of New Territories occupied by agricultural zones and across urban Yuen Long Town before flowing into Shenzhen Bay (Fig. 1). It is ~60 km long, and its catchment basin covers ~27 $km^2$. As far back as the 1990s, pollutants from the river flowing to its mouth at Shenzhen Bay have contaminated the waters of the bay (Qiu 1999). As of 2014, it receives the poorest water quality index grading among all rivers in Hong Kong. Pollutant levels tend to peak during northern hemisphere winter months when discharge and rainfall are low (Qiu 1999). The major pollution sources are unregulated discharges from livestock farms and untreated sewage from urban areas within the basin (Environmental Protection Department of Hong Kong 2007). The Environmental Protection Department of Hong Kong (EPDHK) and the government implemented various laws and built sewage infrastructures for villages to control the pollutant loadings of Hong Kong's rivers and to improve long-term water quality (Environmental Protection Department of Hong Kong 2014). Of the BOD time series at four monitoring points in Yuen Long Creek, all but YL4 recorded improved BODs after 2008 (see the Appendix).

### Material

The water quality data considered in this study were collected by the EPDHK once per month at four monitoring points (YL1, YL2, YL3, and YL4; Fig. 1), situated at tributaries of the Yuen Long Creek, during the period extending from January 2000 to December 2014. The Yuen Long Creek is part of the Deep Bay water control zone at the northwestern part of New Territories. Forty-one water quality variables including BOD are available (Table 1). According to the EPDHK (2014), the physical/chemical properties (temperature, conductivity, flow, etc.) were measured on site using a water quality data logger and an electromagnetic flowmeter. The rest of the water quality parameters related to pollutants (solids, aggregate organics, bacteria, nutrients, and metals) were measured using various in-house pollutant measurement equipment

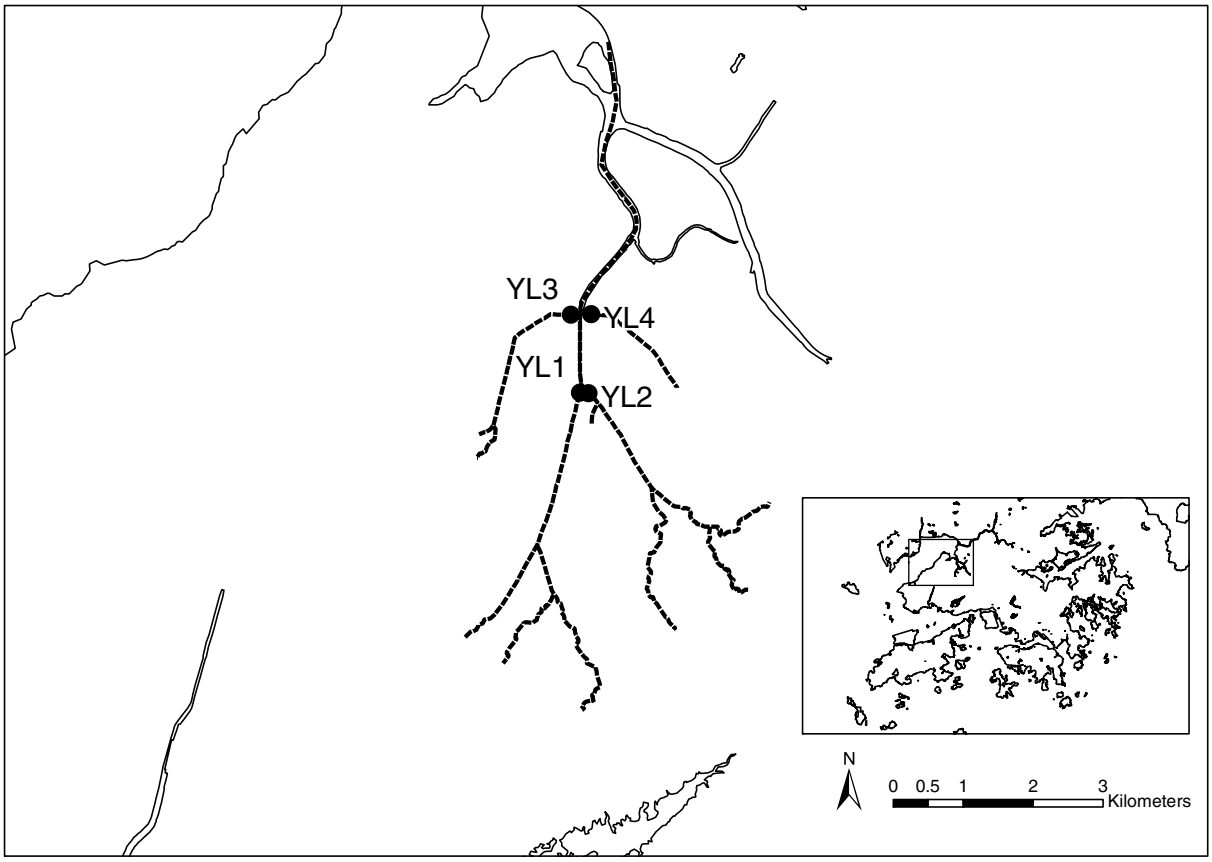


**Fig. 1.** Yuen Long Creek basin; inset map shows the basin's location in Hong Kong, with the river indicated by a dashed line and the four monitoring points indicated by black dots

**Table 1.** Means, Standard Deviations (SD), and Ranges of Water Quality Variables for the Four Yuen Long Creek Tributaries

| | YL1 ($n = 166$) | | YL2 ($n = 170$) | | YL3 ($n = 160$) | | YL4 ($n = 162$) | |
|---|---|---|---|---|---|---|---|---|
| Variable | Mean ± SD | Range | Mean ± SD | Range | Mean ± SD | Range | Mean ± SD | Range |
| 5-day BOD ($mgL^{-1}$) | 35 ± 47 | 1–300 | 12 ± 16 | 1–120 | 57 ± 44 | 3–350 | 91 ± 57 | 4–500 |
| Salinity (psu) | 0.21 ± 0.12 | 0.1–0.7 | 0.19 ± 0.05 | 0.1–0.4 | 0.28 ± 0.46 | 0.1–6 | 0.39 ± 0.58 | 0.0–5.9 |
| Conductivity ($\mu Scm^{-1}$ ) | 440 ± 220 | 110–1,400 | 430 ± 110 | 110–910 | 570 ± 850 | 130–11,000 | 760 ± 1, 100 | 94–11,000 |
| pH | 7.5 ± 0.3 | 6.7–9.2 | 7.5 ± 0.2 | 6.9–8.3 | 7.5 ± 0.2 | 6.8–8.5 | 7.4 ± 0.4 | 6.8–9.8 |
| Water temperature (°C) | 25 ± 4 | 15–33 | 27 ± 4 | 15–35 | 25 ± 4 | 15–33 | 25 ± 4 | 15–32 |
| Flow ($m^3 s^1$) | 0.26 ± 0.76 | 0.0–7.0 | 0.13 ± 0.26 | 0.01–2.9 | 0.67 ± 0.66 | 0.01–5.4 | 0.24 ± 0.25 | 0.0–1.9 |
| Dissolved oxygen ($mgL^{-1}$) | 5.3 ± 1.8 | 1.8–9.7 | 6.9 ± 1.6 | 2–11 | 4.0 ± 1.5 | 1.2–7.8 | 4.2 ± 1.5 | 1.4–9.0 |
| Turbidity (NTU) | 40.9 ± 95 | 3–1,100 | 31 ± 92 | 2–1,000 | 49 ± 60 | 5–51 | 66 ± 120 | 5–1,000 |
| Total solids ($mgL^{-1}$) | 300 ± 190 | 110–1,700 | 270 ± 80 | 130–850 | 390 ± 520 | 100–6,600 | 620 ± 770 | 160–7,000 |
| Total suspended solids ($mgL^{-1}$) | 45 ± 75 | 1–540 | 22 ± 42 | 1–430 | 68 ± 99 | 3–980 | 87 ± 130 | 5–1,200 |
| Total volatile solids ($mgL^{-1}$) | 96 ± 72 | 14–460 | 76 ± 36 | 12–300 | 120 ± 100 | 21–1,200 | 190 ± 200 | 13–1,500 |
| Total dissolved nonvolatile solids ($mgL^{-1}$) | 160 ± 98 | 0–940 | 170 ± 43 | 62–400 | 210 ± 420 | 12–5,400 | 340 ± 560 | 20–5,500 |
| Total phosphorus ($mgL^{-1}$) | 2.9 ± 3.0 | 0–15 | 2.4 ± 1.0 | 0.3–5.8 | 2.8 ± 2.4 | 0–12 | 1.5 ± 0.6 | 0.3–5.1 |
| Orthophosphate ($mgL^{-1}$) | 2.2 ± 2.3 | 0–11 | 2.1 ± 0.9 | 0.2–5.1 | 1.9 ± 1.6 | 0.0–6.9 | 0.73 ± 0.44 | 0.0–3.8 |
| Other phosphorus ($mgL^{-1}$) | 0.7 ± 1.0 | 0.0–8.0 | 0.38 ± 0.48 | 0.0–4.8 | 0.92 ± 0.88 | 0.1–5.3 | 0.78 ± 0.36 | 0.1–2.7 |
| Total organic carbon ($mgL^{-1}$) | 20 ± 30 | 2–180 | 11 ± 14 | 2–140 | 23 ± 22 | 3–150 | 28 ± 15 | 2–94 |
| Total Kjeldahl nitrogen ($mgL^{-1}$) | 20 ± 23 | 1–120 | 12 ± 6 | 1–35 | 17 ± 13 | 1–61 | 11 ± 4 | 1–33 |
| Ammonia ($mgL^{-1}$) | 16 ± 20 | 0–110 | 10 ± 5 | 0–31 | 12 ± 10 | 0–44 | 6.5 ± 3.0 | 0–24 |
| Nirate ($mgL^{-1}$) | 0.63 ± 0.56 | 0.0–3.2 | 2.2 ± 1.8 | 0–12 | 0.30 ± 0.53 | 0.0–2.6 | 0.16 ± 0.52 | 0.0–4.6 |
| Non-ammonia nitrogen | 3.6 ± 4.8 | 0–24 | 2.0 ± 1.6 | 0–10 | 4.6 ± 4.3 | 0–22 | 4.6 ± 1.9 | 1–10 |
| Molybdate-reactive silica ($mgL^{-1}$) | 15 ± 2 | 3–22 | 14 ± 2 | 5–18 | 15 ± 2 | 3–19 | 14 ± 3 | 2–29 |
| Oil and grease ($mgL^{-1}$) | 2.7 ± 6.2 | 0–37 | 0.8 ± 1.0 | 0–10 | 7 ± 11 | 0–71 | 9.2 ± 8.9 | 0–45 |
| Boron ($\mu gL^{-1}$) | 35 ± 21 | 25–120 | 37 ± 19 | 25–125 | 43 ± 62 | 25–785 | 53 ± 68 | 25–778 |
| Chloride ($mgL^{-1}$) | 34 ± 32 | 5–260 | 29 ± 15 | 5–170 | 70 ± 250 | 5–3,200 | 150 ± 330 | 5–3,200 |
| Fluoride ($mgL^{-1}$) | 0.36 ± 0.09 | 0.10–0.60 | 0.40 ± 0.12 | 0.10–0.80 | 0.46 ± 0.10 | 0.10–0.80 | 0.52 ± 0.11 | 0.10–1.10 |
| Sulphide ($mgL^{-1}$) | 0.06 ± 0.11 | 0.01–0.75 | 0.02 ± 0.02 | 0.01–0.17 | 0.07 ± 0.08 | 0.01–0.45 | 0.07 ± 0.07 | 0.01–0.33 |
| Fecal coliforms (cfu per 100 mL) | $1.3 \times 10^6 \pm 2.7 \times 10^6$ | $2.8 \times 10^4$–$2.5 \times 10^7$ | $2.7 \times 10^5 \pm 5.0 \times 10^6$ | 5–$5.9 \times 10^6$ | $2.4 \times 10^6 \pm 2.1 \times 10^7$ | $2.4 \times 10^5$–$1.3 \times 10^7$ | $4.3 \times 10^6 \pm 3.5 \times 10^7$ | 0.5–$2.3 \times 10^7$ |
| E. coli (cfu per 100 mL) | $9.7 \times 10^5 \pm 2.0 \times 10^7$ | $2.3 \times 10^4$–$1.8 \times 10^7$ | $1.5 \times 10^5 \pm 4.4 \times 10^6$ | $4 - 5.5 \times 10^6$ | $1.2 \times 10^6 \pm 1.4 \times 10^7$ | $7.1 \times 10^4$–$8.0 \times 10^6$ | $1.6 \times 10^6 \pm 1.1 \times 10^7$ | 0.5–$5.7 \times 10^6$ |
| Aluminum ($\mu gL^{-1}$) | 320 ± 360 | 25–3,600 | 250 ± 430 | 25–5,100 | 320 ± 300 | 25–2,900 | 420 ± 500 | 80–3,900 |
| Anionic surfactants ($mgL^{-1}$) | 0.35 ± 0.27 | 0.0–1.4 | 0.4 ± 0.4 | 0.0–1.7 | 1.2 ± 0.6 | 0.0–3.1 | 2.1 ± 1.0 | 0.0–5.6 |
| Arsenic ($\mu gL^{-1}$) | 2.5 ± 4.2 | 1–39 | 1.9 ± 1.4 | 1–16 | 2.5 ± 1.6 | 1–9 | 3.0 ± 1.7 | 1–13 |
| Barium ($\mu gL^{-1}$) | 23 ± 12 | 2–110 | 25 ± 17 | 12–210 | 30 ± 17 | 13–110 | 40 ± 26 | 14–180 |
| Cadmium ($\mu gL^{-1}$) | 0.16 ± 0.25 | 0.1–1.9 | 0.1 ± 0.1 | 0.1–1.4 | 0.18 ± 0.20 | 0.1–1.3 | 0.15 ± 0.17 | 0.1–1.2 |
| Chromium ($\mu gL^{-1}$) | 1.1 ± 1.5 | 1–12 | 0.7 ± 1.1 | 1–13 | 1.5 ± 1.6 | 1–12 | 1.6 ± 2.4 | 1–19 |
| Copper ($\mu gL^{-1}$) | 18 ± 18 | 0–120 | 7.4 ± 8.2 | 1–92 | 22 ± 23 | 2–130 | 9.0 ± 9.6 | 2–77 |
| Iron ($\mu gL^{-1}$) | 640 ± 570 | 25–7,100 | 610 ± 810 | 150–8,200 | 710 ± 550 | 190–5,900 | 740 ± 650 | 320–6,300 |
| Lead ($\mu gL^{-1}$) | 11 ± 29 | 1–280 | 4 ± 11 | 1–130 | 9 ± 14 | 1–110 | 8 ± 13 | 1–100 |
| Manganese ($\mu gL^{-1}$) | 180 ± 93 | 5–700 | 130 ± 83 | 26–810 | 190 ± 110 | 5–960 | 180 ± 100 | 60–820 |
| Nickel ($\mu gL^{-1}$) | 3.0 ± 2.2 | 1–16 | 3.0 ± 1.5 | 1–17 | 4.0 ± 2.3 | 1–18 | 4.6 ± 2.23 | 1–16 |
| Zinc ($\mu gL^{-1}$) | 95 ± 140 | 5–1,400 | 51 ± 110 | 15–1,400 | 100 ± 100 | 18–830 | 66 ± 54 | 20–420 |

Note: $n$ = number of instances. All values are rounded up to two significant figures.

in the EPDHK laboratory. Further details on the equipment used can be found in EPDHK (2014). In this paper, BOD is regressed against 40 other variables.

Site YL2 has the lowest mean BOD (12 mg $L^{-1}$) of all monitoring points (Table 1). Site YL4 has the highest mean (91 mg $L^{-1}$) and the largest range (5–500 mg $L^{-1}$). Sites YL1 and YL3 have intermediate means (35 and 57 mg $L^{-1}$ respectively) and relatively large ranges on the order of <5 to ≥ 300 mg $L^{-1}$. Despite some differences in mean BOD, the data from all monitoring stations are combined in this analysis to form one large data set with 658 instances. The rationale for combination is to create more data for training of the models in order to reduce model overfitting.

The missing and inaccurate BODs at all monitoring points are mainly found in the years 2000–2005 (see the Appendix); these are the values this paper attempts to fill. The absence of missing and inaccurate BODs after 2005 perhaps suggests a recent improvement in BOD testing procedures that mitigates the problem of erroneous BOD test results. Nevertheless, it is important to correct the older data to facilitate studies investigating water quality changes over time.

### Support Vector Regression

This section briefly derives the SVR function, which is developed by Vapnik (1995). Detailed mathematical treatment of SVR can be found in the literature by Vapnik (1995) and Smola and Scholkopf (2004). Basically, the SVR maps the training set $\{x_1, O_1), \ldots, (x_N, O_N)\} \subset \mathcal{X} \times \mathbb{R}$ nonlinearly from the original input space $\mathcal{X}$ to a higher dimensional feature space $\mathcal{F}$ in order to perform the linear separation that solves the regression problem. The SVR function is given by the following:

$$f(x) = w \cdot \varphi(x) + b \quad \text{with} \quad w \in \mathcal{X}, b \in \mathbb{R} \tag{1}$$

where $\varphi(x)$ = nonlinear mapping function that transforms the input data from $\mathcal{X}$ to $\mathcal{F}$; $w$ = weight vector; and $b$ = bias term. The aim is to establish Eq. (1) such that the linear separation can tolerate up to a fixed error $\varepsilon$ while being as flat as possible such that $\|w\|$ is minimized. With the introduction of slack variables $\xi$ and $\xi^*$, this can be seen as the convex optimization problem:

$$\text{minimize} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$$
$$\text{subject to} \begin{cases} O_i - f(x_i) & \leqslant \varepsilon + \xi_i \\ f(x_i) - O_i & \leqslant \varepsilon + \xi_i^* \\ C & > 0 \\ \xi_i, \xi_i^* & \geqslant 0 \end{cases} \tag{2}$$

where $C$ = trade-off between the model flatness and the amount up to which errors larger than $\varepsilon$ are tolerated.

It follows that $\xi_i$ and $\xi_i{}^*$ are zero only when $O_i - f(x_i) \leqslant \varepsilon$ and $f(x_i) - O_i \leqslant \varepsilon$, respectively.

The optimization problem can be solved more easily in its dual optimization form by introducing a dual set of Lagrange multipliers $(\alpha, \alpha^*, \eta, \eta^*)$ to form the Lagrangian $L$:

$$L = \begin{cases} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) - C\sum_{i=1}^{N}(\eta_i\xi_i + \eta_i\xi_i^*) \\ -\sum_{i=1}^{N}\alpha_i[\varepsilon + \xi_i - O_i + f(x_i)] \\ -\sum_{i=1}^{N}\alpha_i^*[\varepsilon + \xi_i^* + O_i - f(x_i)] \end{cases}$$
$$\text{subject to } \alpha, \alpha^*, \eta, \eta^* \geqslant 0 \tag{3}$$

Optimization is the minimization of $L$ with respect to the primal variables $(w, \xi_i, \xi_i^*)$ and simultaneous maximization with respect to the Lagrange multipliers. This implies that the solution has a saddle point, which means that the partial derivatives of $L$ with respect to the primal variables must be zero for optimality. Substituting these partial derivatives of $L$ in Eq. (3) yields the reformulated dual optimization problem:

$$\text{maximize} \begin{cases} \sum_{i=1}^{N} O_i(\alpha_i - \alpha_i^*) - \varepsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) \\ -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\varphi(x_i)\cdot\varphi(x_j) \end{cases}$$
$$\text{subject to} \begin{cases} \sum_{i=1}^{N}(\alpha - \alpha^*) = 0 \\ \alpha, \alpha^* \in [0, C] \end{cases} \tag{4}$$

Solving the optimization problem allows Eq. (1) to be rewritten as

$$f(x) = \sum_{i=1}^{N}(\alpha - \alpha^*)\varphi(x_i)\cdot\varphi(x) + b \tag{5}$$

The final step is to deal with the unknown $\varphi(x)$. Instead of finding a suitable $\varphi(x)$, because Eqs. (4) and (5) depend only on the dot products between $x$, a kernel function $k(x_i, x_j) = \varphi(x_i)\cdot\varphi(x_j)$ can be employed for computational efficiency. Hence, the final form of the SVR function is

$$f(x) = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)k(x_i, x) + b \tag{6}$$

The value of $b$ can be calculated by applying the Karush–Kuhn–Tucker conditions to the dual optimization problem, during which, at optimality, the constraints and products between dual variables are zero (Karush 1939; Kuhn and Tucker 2014).

Each $(x_i, O_i)$ whose corresponding $\alpha_i$ and $\alpha_i^*$ are nonzero and contribute to the SVR function are known as support vectors (SVs). There are various kernel functions, but this paper adopts the radial basis function, which is the most commonly used kernel function and is the best at modeling nonlinear relations (Hsu et al. 2003; Liu and Lu 2014). The radial basis function $k_{\text{RBF}}$ is defined as

$$k_{\text{RBF}}(x_i, x) = e^{-\gamma|x_i - x^2|} \tag{7}$$

where $\gamma$ = free parameter.

Hyperparameter optimization is conducted to find the optimal parameters, $\gamma$, $C$ and $\varepsilon$, via a 10-fold cross validation exhaustive grid search in $R$. The procedure generates the model based on all possible combinations of $\gamma$, $C$, and $\varepsilon$, where $\gamma \in [2^{-5}, 2^{-1}]$, $C \in [1, 16]$, and $\varepsilon \in [0, 1]$, in order to find the best combination that produces the least model error. All training and testing sets are normalized to the range (0,1) before building the SVR model using $R$ in order for all variables, regardless of their magnitudes, to be weighted equally (Hsu et al. 2003).

### Model Performance

The conventional way to evaluate model performance is to randomly split the data set into separate training, validation, and testing sets. The training set is for fitting the model parameters; the validation set is for determining the optimal model parameters; and the testing set is for assessing generalization errors on the finalized model using unfamiliar instances (Hastie et al. 2009). Because the goal in this paper is only to assess how changing the proportion

of extreme-high BODs in a training set affects SVR model performance, only the training and testing phases are carried out.

In order to define the critical cutoff for extreme-high BODs—that is, those subject to systematic underprediction with consistently negative error residuals—first a 10-fold cross validation is run in which the full data set is randomly split into 10 equal subsamples [according to Kohavi (1995), 10 is the optimal number of folds for a dataset size on the order of hundreds to obtain the residual errors of prediction outputs]. For the first cycle, a single subsample is selected as the testing set and the remaining nine form the training set to evaluate testing performance. For the next 9 cycles, a different subsample is selected as the testing set and the rest make up the training set. The 10 results are then averaged to produce a single estimation of model performance using a variety of metrics (as described subsequently).

To identify the critical cutoff BOD separating the extreme-high values from the rest, all residual errors obtained from the 10-fold cross validation are arranged in ascending order of BOD. Positive (negative) residual errors are then denoted $+1(-1)$. The assigned values can be separated into two clusters, one for the lower range of BODs and the other for the upper range, and the mean assigned values for each cluster are calculated. The cutoff value is identified by finding the cluster size that produces the largest mean difference between the two clusters.

The analysis begins by splitting the full data set randomly, with ~80% (528 instances) of the values reserved for training and ~20% (130 instances) reserved for testing. The proportion of extreme-high BODs in both training and testing data sets is approximately the same as that in the full data set, with the proportion of extreme-high values accounting for 3.6% of the full data set, 3.0% of the training set, and 6.1% of the testing set. To test the hypothesis that raising the proportion of extremely high values in the training set increases their weighting in the model, which in turn mitigates systematic underprediction, the training set is modified by increasing the proportion of extreme-high values in the training set. Two novel methods are used to evaluate this approach. The first method, bootstrapping, first randomly selects and deletes BODs below the threshold value of 150 mg $L^{-1}$ before duplicating an equal number of BODs above the threshold value to maintain the size of the training set while increasing the proportion of extreme-high values. Duplication is repeated to increase the proportion of extreme-high values to ~10, ~20, ~30, ~40, and ~50% of the training set. For each proportion of extreme-high values, bootstrapping is carried out 10 times to form 10 new random data sets. Ten models are built for each training set, and the average performance of all 10 is then evaluated.

The second method is simple data expansion, for which BODs above the threshold value are duplicated and then added back into the training set without deleting nonextreme values. Consequently, the size of the training set increases. With this approach, five new data sets are created for which the proportions of extreme-high values are ~10, ~20, ~30, ~40, and ~50% of the total size of the data set.

Despite the different training sets used to build the models, hyperparameter optimization is conducted only once on the entire data set of 658 instances and the same set of obtained optimal parameters is applied to build all models. This is because, first, the focus is on evaluating the effectiveness of both methods in mitigating systematic underprediction for extreme-high values in general, not on building the best model for each training set. Second, the focus is not determining which method is better but whether both methods work. Hence, it is sufficient to use the same set of optimal parameters to build approximately optimal models.

All models built for these 10 new data sets are evaluated with four criteria: (1) mean relative absolute error (MAE), (2) bias, (3) coefficient of determination ($R^2$), and (4) Nash–Sutcliffe efficiency (NSE). The MAE measures the mean of the magnitudes of all residual errors between the modeled ($y_i$) and observed ($O_i$) values of the dependent variable (BOD). Bias represents the mean of all residual errors (Liu and Lu 2014), indicating the extent to which the modeled values overestimate or underestimate the dependent variable. The $R^2$ (of regression) measures the percentage of data variability explained by the modeled values and the goodness of fit, ranging 0 to 1 (zero correlation to a perfect match between the modeled and observed values). The NSE (Nash and Sutcliffe 1970) is a variance-normalized statistic measuring how close the modeled and observed values are. An NSE value of 1 corresponds to a perfect match between the modeled and observed values; a value of 0 indicates that the model predictions are as accurate as the mean of observed values. A negative NSE value indicates that the model is predicting worse than than if the mean of observed values were used. The mathematical formulas for the four assessment metrics are not reproduced in this paper because they can be found in the modeling literature (e.g., Liu and Lu 2014).

## Results

### Systematic Underprediction and Critical Cutoff BOD of Extreme-High Values

The 10-fold cross validation grid search on the full data set of 658 instances yields the following optimal model parameters: $\gamma = 0.03$, $C = 4$, and $\varepsilon = 0.04$. Using the previously described method, the critical cutoff BOD is identified to be 150 mg $L^{-1}$. From Table 2 and Fig. 2, although the high $R^2$ of 0.68 and positive NSE of 0.66 for the full data set indicate that the SVR's performance is adequate, there are modeled values falling well below the observed values at the extreme-high end of the range. Such underprediction does not affect nonextreme values, as is evident from the positive bias, high $R^2$, and positive NSE for values <150 mg $L^{-1}$. This pattern

**Table 2.** Statistics of 10-Fold Cross Validation of SVR Models Using the Full Data Set of 658 Values

| Statistic | All BOD | Nonextreme-high BOD (<150 mg $L^{-1}$) | Extreme-high BOD (≥150 mg $L^{-1}$) |
|---|---|---|---|
| MAE | 16.90 | 13.91 | 96.11 |
| Bias | 0.66 | 4.33 | −96.11 |
| $R^2$ | 0.66 | 0.76 | 0.08 |
| NSE | 0.66 | 0.74 | −1.54 |

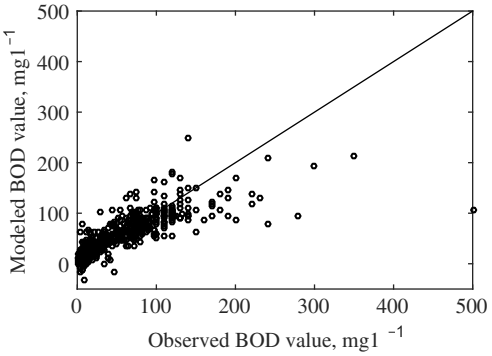


**Fig. 2.** Plot of observed versus SVR-modeled BODs from 10-fold cross validation using the full data set of 658 values

represents a systematic underprediction for extreme-high values at a threshold value of approximately 150 mg $L^{-1}$.

Only 3.6% of the BODs in the full data set are greater than the 150-mg $L^{-1}$ threshold value. Ranging 150–500 mg $L^{-1}$, they cover 70% of the entire range of observed values (1–500 mg $L^{-1}$).

### Model Characteristics

For each model, the instances used in both training and testing sets and those belonging to extreme-high BODs are tabulated in Table 3. Table 4 shows the number and distribution of SVs and the sums of SVs $\alpha$ and $\alpha^*$ ($\sum \alpha$ and $\sum \alpha^*$) falling in the nonextreme and extreme-high ranges. The frequency distributions of SVs for all models are plotted in Fig. 3. For the bootstrapping approach, for each proportion of extreme-high values in the training set, the SVs $\alpha$ and $\alpha^*$ from all 10 models are added together (Table 4). The majority of SVs with nonzero $\alpha$ ($\alpha^*$) correspond to the extreme-high (nonextreme) values.

### Model Performance

According to Table 5, with an increasing proportion of extreme-high BODs in the training set, the bootstrapping and data expansion methods lead to a less negative bias in model performance on training sets for all BODs but to an increase in MAE. Other metrics such as $R^2$ and NSE remain relatively constant. This is associated with a more positive bias for nonextreme values and a less negative bias in performance for extreme-high values.

The final evaluation of the two methods is based on performance in the testing phase as shown in Table 6 and Fig. 4. Model performance for the extreme-high values generally increase as the proportion of high values increases from 3 to ~30–40%, where the systematic underprediction problem for extreme-high values is resolved, although there are some disagreements among metrics. For the bootstrapping approach, $R^2$ for extreme-high values peaks at 0.28 when the proportion of high values is ~20%. Bias and NSE improve by approaching zero at ~40%. Also, the MAE continues to decrease until the proportion of high values reaches ~40%. Almost identical results are observed for the data expansion method.

Such improvement in the prediction of extreme-high values occurs at the slight expense of accurate prediction of nonextreme values (<150 mg $L^{-1}$). For both bootstrapping and data expansion, $R^2$ for nonextreme values decreases from 0.69 to 0.65 when the proportion of extreme-high values is adjusted to ~20%. In addition, the corresponding MAE and bias increased and NSE turned negative as the proportion of high values increased.

**Table 3.** BOD Statistics in Training and Testing Sets Generated by the Investigated Bootstrapping and Data Expansion Methods, with the Percentage of Extreme-High BODs Being in the Original Dataset Being 3%

| Extreme-high BODs in training set (approximate %) | Instances in training set | Extreme-high BODs in training set (≥150 mg $L^{-1}$) | Instances in testing set | Extreme-high BODs in testing set (≥150 mg $L^{-1}$) |
|---|---|---|---|---|
| 3 | 528 | 16 | 130 | 8 |
| | | Bootstrapping | | |
| 9.1 (~10) | 528 | 48 | 130 | 8 |
| 21.2 (~20) | 528 | 112 | 130 | 8 |
| 30.3 (~30) | 528 | 160 | 130 | 8 |
| 39.4 (~40) | 528 | 208 | 130 | 8 |
| 51.5 (~50) | 528 | 272 | 130 | 8 |
| | | Data expansion | | |
| 11.1 (~10) | 576 | 64 | 130 | 8 |
| 20.0 (~20) | 640 | 128 | 130 | 8 |
| 30.4 (~30) | 736 | 224 | 130 | 8 |
| 39.6 (~40) | 848 | 336 | 130 | 8 |
| 50.0 (~50) | 1024 | 512 | 130 | 8 |

**Table 4.** SV Statistics in Nonextreme-High and Extreme-High BOD Range across Percentages of Extreme-High BODs in the Training Set for the Investigated Methods

| | Nonextreme-high BODs (≥150 mg $L^{-1}$) | | | | Extreme-high BODs (≥150 mg $L^{-1}$) | | | |
|---|---|---|---|---|---|---|---|---|
| Percentage | $\sum \alpha$ | $\sum \alpha^*$ | SVs with nonzero $\alpha$ | SVs with nonzero $\alpha^*$ | $\sum \alpha$ | $\sum \alpha^*$ | SVs with nonzero $\alpha$ | SVs with nonzero $\alpha^*$ |
| 3 | 146 | 208 | 44 | 61 | 62 | 0 | 16 | 0 |
| | | | | Bootstrapping | | | | |
| ~10 | 1,106 | 2,574 | 343 | 723 | 1,468 | 0 | 389 | 0 |
| ~20 | 858 | 3,342 | 295 | 919 | 2,484 | 0 | 647 | 0 |
| ~30 | 795 | 3,727 | 268 | 1,005 | 3,017 | 85 | 776 | 32 |
| ~40 | 718 | 3,877 | 239 | 1,034 | 3,495 | 336 | 897 | 99 |
| ~50 | 759 | 4,043 | 234 | 1,067 | 4,029 | 744 | 1,038 | 211 |
| | | | | Data expansion | | | | |
| ~10 | 106 | 289 | 33 | 79 | 183 | 0 | 48 | 0 |
| ~20 | 103 | 393 | 33 | 110 | 290 | 0 | 75 | 0 |
| ~30 | 110 | 502 | 36 | 135 | 402 | 10 | 103 | 4 |
| ~40 | 130 | 620 | 41 | 161 | 541 | 50 | 137 | 14 |
| ~50 | 183 | 797 | 51 | 207 | 734 | 120 | 192 | 35 |



J. Environ. Eng., 2017, 143(9): -1--1

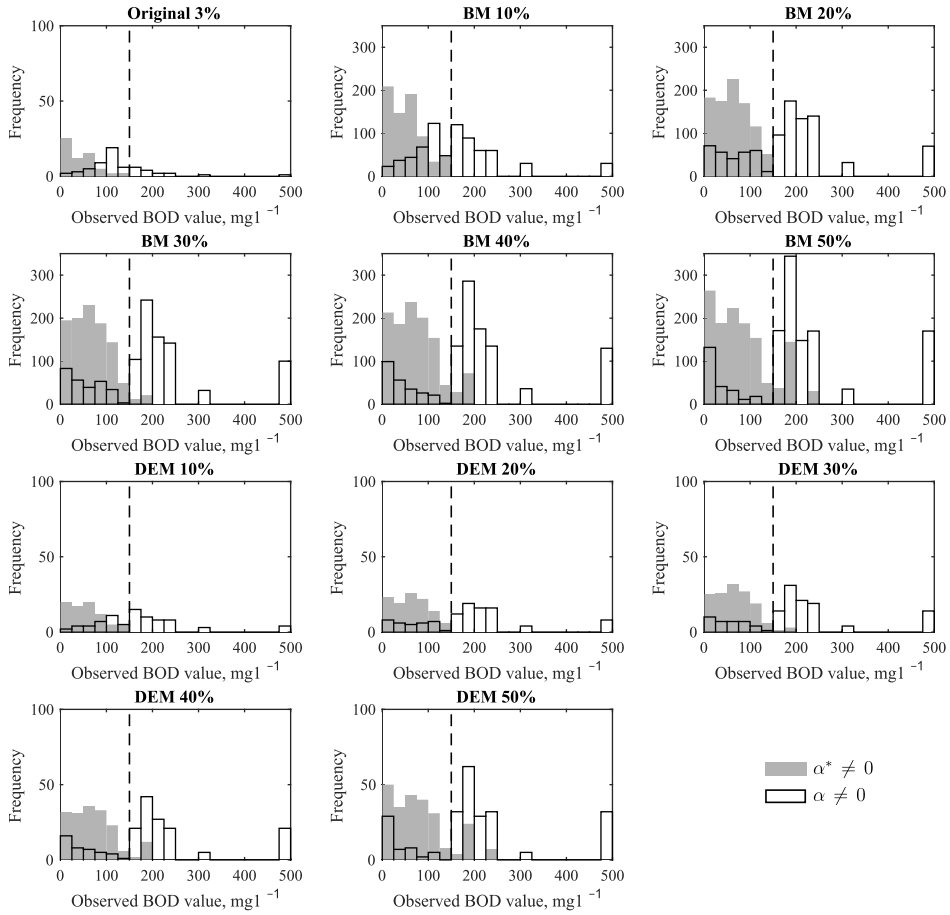


**Fig. 3.** Frequency distributions of model SVs generated across different percentages of extreme-high BODs in the training set using the investigated methods; the distribution of SVs with $\alpha \neq 0$ ($\alpha^* \neq 0$) is indicated by gray bars; observed BOD of 150 mg $L^{-1}$ is indicated by dashed lines

**Table 5.** Training-Phase Model Performances for All BODs, Nonextreme-High BODs, and Extreme-High BODs across Percentages of Extreme-High BODs in the Training Set Generated by the Investigated Bootstrapping and Data Expansion Methods, with the Percentage of Extreme-High BODs Being in the Original Data Set Being 3%

| | All BODs | | | | Nonextreme-high BODs (≥150 mg $L^{-1}$) | | | | Extreme-high BODs (≥150 mg $L^{-1}$) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Percentage | MAE | Bias | $R^2$ | NSE | MAE | Bias | $R^2$ | NSE | MAE | Bias | $R^2$ | NSE |
| 3 | 14.73 | 1.02 | 0.71 | 0.70 | 12.30 | 3.94 | 0.83 | 0.82 | 92.50 | −92.50 | 0.01 | −1.31 |
| | | | | | | Bootstrapping | | | | | | |
| ~10 | 18.46 | −1.88 | 0.66 | 0.66 | 13.02 | 5.18 | 0.81 | 0.77 | 72.83 | −72.42 | 0.02 | −0.93 |
| ~20 | 24.44 | −5.68 | 0.68 | 0.67 | 16.06 | 7.26 | 0.78 | 0.60 | 55.55 | −53.71 | 0.07 | −0.41 |
| ~30 | 28.16 | −5.92 | 0.68 | 0.68 | 19.00 | 9.98 | 0.77 | 0.45 | 49.23 | −42.50 | 0.11 | −0.19 |
| ~40 | 30.81 | −6.25 | 0.69 | 0.68 | 22.01 | 12.70 | 0.77 | 0.27 | 44.35 | −35.41 | 0.17 | −0.01 |
| ~50 | 33.13 | −2.64 | 0.72 | 0.71 | 29.21 | 19.88 | 0.73 | −0.26 | 36.82 | −23.83 | 0.42 | 0.31 |
| | | | | | | Data expansion | | | | | | |
| ~10 | 19.32 | −2.90 | 0.67 | 0.66 | 13.15 | 5.32 | 0.81 | 0.75 | 68.67 | −68.67 | 0.02 | −0.81 |
| ~20 | 23.80 | −5.19 | 0.68 | 0.67 | 15.83 | 7.07 | 0.79 | 0.62 | 55.69 | −54.23 | 0.07 | −0.42 |
| ~30 | 27.81 | −6.24 | 0.69 | 0.68 | 18.72 | 9.53 | 0.77 | 0.47 | 48.61 | −42.27 | 0.10 | −0.18 |
| ~40 | 30.50 | −6.18 | 0.69 | 0.68 | 22.04 | 12.54 | 0.76 | 0.26 | 43.39 | −34.70 | 0.18 | 0.00 |
| ~50 | 32.70 | −1.48 | 0.73 | 0.72 | 29.03 | 18.76 | 0.71 | −0.24 | 36.37 | −21.72 | 0.43 | 0.33 |

In terms of $R^2$, the effect of adjusting the proportion of extreme-high values on testing performance for all BODs is not large for either method. For all BODs, $R^2$ decreases slightly (0.71 to 0.68) as the proportion of high values reaches ~20%. Substantial changes are observed for other metrics, where the corresponding MAE and bias increase and the NSE decreases as the extreme-high value proportion increases to 30%. Despite the changes in MAE and bias, indicating worsening model performance, the high $R^2$ and positive NSE of >0.3 for the proportion of ~3–40% illustrates that the model's predictive capability for all BODs is still robust.

For both methods, the best performances (in terms of MAE and bias) in predicting extreme-high values results from increasing the proportion of extreme-high values in the training set to ~30–40%. Increases above ~40% cause model performance to decrease.

**Table 6.** Testing-Phase Model Performances for All BODs, Nonextreme-High BODs, and Extreme-High BODs across Percentages of Extreme-High BODs in the Training Set Generated by the Investigated Bootstrapping and Data Expansion Methods, with the Percentage of Extreme-High BODs Being in the Original Data Set Being 3%

| | All BODs | | | | Nonextreme-high BODs (<150 mg $L^{-1}$) | | | | Extreme-high BODs (≥150 mg $L^{-1}$) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Percentage | MAE | Bias | $R^2$ | NSE | MAE | Bias | $R^2$ | NSE | MAE | Bias | $R^2$ | NSE |
| 3 | 19.91 | 0.73 | 0.71 | 0.71 | 16.67 | 5.30 | 0.69 | 0.63 | 69.27 | −69.01 | 0.24 | −1.12 |
| | | | | | Bootstrapping | | | | | | | |
| ~10 | 20.56 | 5.00 | 0.68 | 0.64 | 18.09 | 7.95 | 0.64 | 0.42 | 58.19 | −39.96 | 0.25 | −0.61 |
| ~20 | 24.96 | 11.17 | 0.68 | 0.49 | 22.81 | 12.84 | 0.65 | 0.02 | 57.78 | −14.42 | 0.28 | −0.38 |
| ~30 | 28.76 | 15.48 | 0.68 | 0.41 | 27.16 | 17.09 | 0.66 | −0.17 | 53.14 | −9.09 | 0.25 | −0.13 |
| ~40 | 32.70 | 19.53 | 0.66 | 0.32 | 31.47 | 21.31 | 0.67 | −0.40 | 51.40 | −7.65 | 0.21 | −0.04 |
| ~50 | 41.26 | 27.98 | 0.62 | 0.04 | 40.49 | 29.76 | 0.65 | −1.03 | 52.97 | 0.77 | 0.12 | −0.04 |
| | | | | | Data expansion | | | | | | | |
| ~10 | 21.23 | 6.12 | 0.68 | 0.61 | 18.81 | 8.78 | 0.64 | 0.34 | 58.15 | −34.47 | 0.25 | −0.52 |
| ~20 | 24.60 | 10.63 | 0.68 | 0.50 | 22.5 | 12.3 | 0.65 | 0.05 | 56.64 | −14.76 | 0.29 | −0.34 |
| ~30 | 28.37 | 14.67 | 0.68 | 0.43 | 26.75 | 16.31 | 0.67 | −0.14 | 53.16 | −10.36 | 0.26 | −0.09 |
| ~40 | 32.97 | 19.08 | 0.67 | 0.34 | 31.74 | 20.96 | 0.68 | −0.36 | 51.64 | −9.53 | 0.21 | 0.00 |
| ~50 | 41.10 | 25.62 | 0.61 | 0.07 | 40.27 | 27.6 | 0.64 | −0.96 | 53.82 | −4.63 | 0.09 | −0.06 |

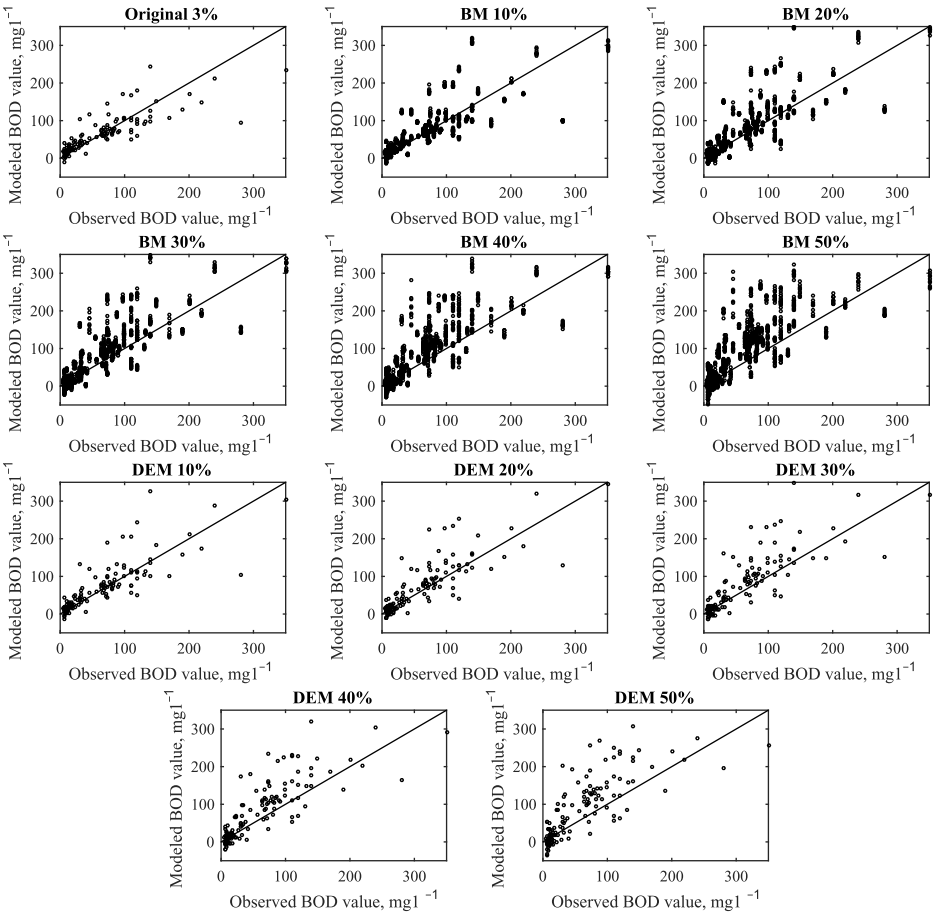


**Fig. 4.** Plots of observed versus modeled BODs based on model performance in the testing phase across percentages of extreme-high BODs in the training set using the investigated methods; the diagonal line is the 1:1 line

## Discussion

### *Analysis of Systematic Underprediction*

Two important questions arising from the results are addressed in this section. The first is why the data sets investigated in this paper are affected by systematic underprediction for extreme-high values. The second question is why systematic underprediction, instead of overprediction, occurs for the extreme-high values.

The answers to both questions lie in the distribution of observed BODs, the SVs in the model generated using the original uncorrected training set, and the kernel function $k_{RBF}$. In the results, a relatively small number of extreme-high BODs (3.6%) are sparsely distributed over the upper 70% of the data range. Moreover, from

Fig. 3(a) and Table 4, it is seen that all SVs in the extreme-high range are associated with nonzero $\alpha$ instead of nonzero $\alpha^*$. The SVR function in Eq. (6) shows that the modeled BOD is the sum of all products between $\alpha$ (or $-\alpha^*$) and $k_{\mathrm{RBF}}$. From Eq. (7), $k_{\mathrm{RBF}}$ is shown to be an exponential function that measures the extent of similarity between two sets of input data $x$ and $x_i$, in which $x_i$ corresponds to a given support vector. When $x$ and $x_i$ are more similar (dissimilar), $k_{\mathrm{RBF}}$ is larger (smaller) (i.e., identical $x$ and $x_i$ lead to $k_{\mathrm{RBF}} = 1$, whereas dissimilar $x$ and $x_i$ lead to $k_{\mathrm{RBF}} \to 0$). It follows that only terms with nonzero $\alpha$ ($\alpha^*$) increase (decrease) the modeled BOD. The variable $k_{\mathrm{RBF}}$ can be seen as giving more (less) weight to $\alpha_i$ or $-\alpha_i^*$ when $x$ is more (less) similar to the $x_i$ that corresponds to $\alpha_i$ or $\alpha_i^*$ for a given $i$. Now consider a given $x$ that corresponds to a extreme-high BOD and is "fed" into the model. Because $x$ is more similar to $x_i$ in SVs with nonzero $\alpha$ lying in the extreme-high range than is $x_i$ of SVs lying in the nonextreme range, and because no SVs with nonzero $\alpha^*$ exist in the extreme-high range, the modeled BOD for that $x$ is generally greater than of other $x$ corresponding to the nonextreme range. However, because the SVs are sparsely distributed throughout the extreme-high range, $k_{\mathrm{RBF}}$ is not sufficiently large to place adequate weights on all SVs falling in the extreme-high range when predicting extreme-high values.

In other words, the sum of products between nonzero $\alpha$ and $k_{\mathrm{RBF}}$ is insufficient to increase a given modeled value (supposedly belonging to the extreme-high range) close to an accurate value, thereby causing the systematic underprediction in this case. This is also to say that systematic overprediction for extreme-high values does not occur for such data sets. In contrast, Fig. 3 and Table 4 indicate that the proportions of SVs associated with nonzero $\alpha$ and $\alpha^*$ falling in the nonextreme range are more balanced and these SVs are densely distributed. Thus, there are sufficient SVs of each type similar enough to a given $x$ corresponding to the nonextreme range such that nonextreme modeled values are not subject to systematic under- or overprediction. Therefore, it can be understood that the SVR algorithmically reduces (increases) the prediction error of dependent variables using $x$ when there are more (fewer) training instances and SVs within a immediate region of input space surrounding $x$.

With respect to studies that reported systematic underprediction affecting peak values (e.g., Balfer and Bajorath 2015; Garsole and Rajurkar 2015; Granata et al. 2016), this study was unable to access their data sets to determine their data distributions and verify the causes of underprediction. Nevertheless, the problem of underprediction affecting extreme-high dependent values is present in the modeling results of these studies, which corroborate the observation here that only underprediction is present in the prediction of extreme-high values.

### Analysis of Model Performance

The results in this study support the hypothesis that more training instances with extreme-high BODs allow the SVR algorithm to mitigate systematic underprediction and reduce the error for extreme-high values, which leads to better model prediction for extreme-high values. As shown by the testing performance results in Table 6, negative bias is reduced, with MAE decreasing up to ~25%. This is accompanied by a trade-off between a lower proportion of nonextreme values that result in greater error tolerance and worsened model prediction for nonextreme values, and a more positive bias and increases in MAE up to ~90%. At the same time, model performance for the entire BOD testing set worsens with a more positive bias and higher MAE because of a greater proportion of nonextreme values than extreme-high values in the testing set.

The explanation for the mitigation of systematic underprediction and error reduction for extreme-high values lies in the Lagrangian problem stated in Eq. (4). From its first term $\sum_{i=1}^{N} O_i(\alpha_i - \alpha_i^*)$, to maximize the function, given that either $\alpha_i$ or $\alpha_i^*$ can be nonzero for a given $i$, most nonzero $\alpha_i$ ($\alpha_i^*$) are associated with large (small) observed values $O_i$. When the proportion of extreme-high values in the training set increases, more SVs with nonzero $\alpha$ falling in the extreme-high range are part of the model. This is consistent with Table 4 and Fig. 3. More SVs with nonzero $\alpha$ can sufficiently increase the modeled value (supposedly belonging to the extreme-high range) close to an accurate value, thereby mitigating systematic underprediction and reducing the error. At the same time, although there are also more SVs with nonzero $\alpha^*$ [following from the constraint $\sum_{i=1}^{N}(\alpha - \alpha^*) = 0$ in Eq. (4)], they mostly fall in the lower end of the nonextreme range, far enough from the extreme-high range not to negate the effect of nonzero $\alpha$ in the extreme-high range (recall the characteristics of $k_{\mathrm{RBF}}$). The explanation is consistent with the observed increase in SVs with nonzero $\alpha$ in the nonextreme range and SVs with nonzero $\alpha^*$ in the extreme-high range, with the increasing proportion of extreme-high values making up the training set (Fig. 3).

It is interesting to observe the more positive bias and greater error for nonextreme values with the proportion of extreme-high values in the training set despite the increase in SVs with nonzero $\alpha^*$ in the nonextreme range. It is observed from both training and testing phases that more frequent overprediction occurs at the upper end (75–150 mg $L^{-1}$) than at the lower end (1–75 mg $L^{-1}$) of the nonextreme range; also, overprediction becomes more frequent with the greater proportion of extreme-high values in the training set (Tables 5 and 6 and Fig. 4). This implies that modeled values in the 75–150 mg $L^{-1}$ range are mainly responsible for the positive bias, the reason being that the majority of SVs with nonzero $\alpha$ lie in the 1–75 mg $L^{-1}$ range and that their nonzero $\alpha$ are given less weight when predicting for those $x$ supposedly corresponding to the 75–150 mg $L^{-1}$ range compared with those $x$ supposedly corresponding to the 1–75 mg $L^{-1}$ range. Coupled with the fact that the 75–150 mg $L^{-1}$ range, rather than the 0.5–75 mg $L^{-1}$ range, is close to the extreme-high range where most SVs with nonzero $\alpha$ are found, overprediction occurs more frequently at the upper end than the lower end of the nonextreme range. Given that systematic underprediction does not occur for the 1–75 mg $L^{-1}$ range even though that is where most nonzero $\alpha^*$ are located, it must follow that overprediction tends to occur in the 75–150 mg $L^{-1}$ range, which consequently results in a positive bias for the entire nonextreme range (1–150 mg $L^{-1}$). With a higher proportion of extreme-high values in the training set and more SVs in the extreme-high range being included in the SVR function, more frequent overprediction in the 75–150 mg $L^{-1}$ range occurs, which leads to a more positive bias and greater error for the nonextreme range. Nevertheless, predictions in the 75–150 mg $L^{-1}$ range are not consistently overestimated and this cannot be classified as systematic overprediction.

In addition, for both methods it is observed that testing performance in terms of MAE for extreme-high values peak when the proportion of extreme-high values increases to ~40%. Conversely, training performance continuously improves with the increasing proportion of extreme-high values (Table 5). These observations indicate that the SVR algorithm overfits for extreme-high values. Having more replicated extreme-high values above a certain threshold for the SVR algorithm to learn leads to improved training performance but poor generalization ability that subsequently worsens testing performance for the entire BOD range (Liu and Lu 2014).

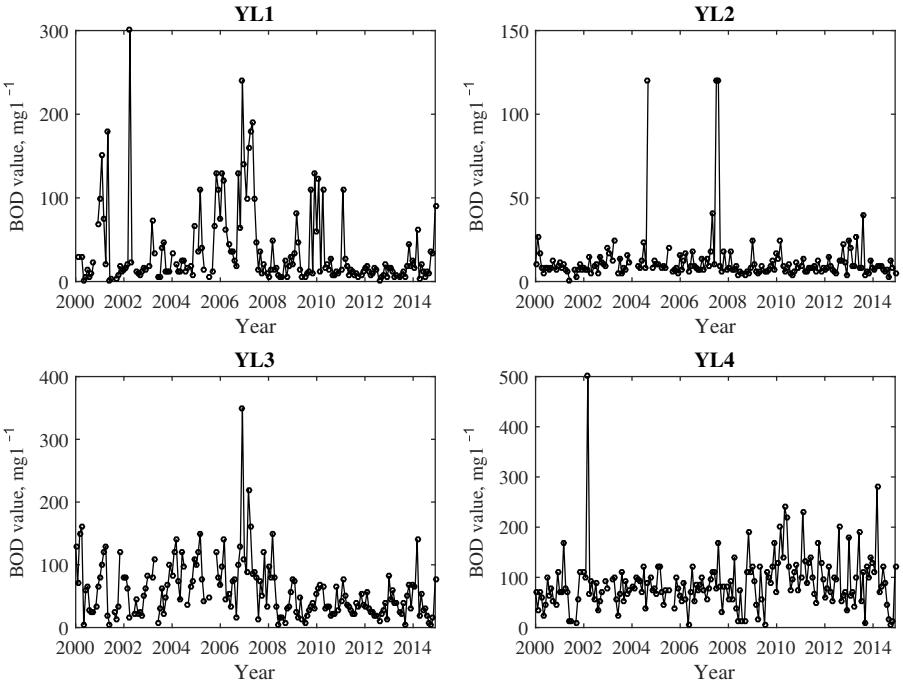


**Fig. 5.** BOD time series for Monitoring Points YL1, YL2, YL3, and YL4 over the period 2000–2014; missing or inaccurate BODs are indicated by gaps in the time series

### *Limitation and Future Research*

One limitation of this study is that, when the relatively small data set sample is partitioned into training and testing sets, there are only eight testing instances of extreme-high BODs in the latter. This small number prevents more accurate interpretations of testing performance, particularly in terms of $R^2$. Further research on larger data sets with more extreme-high values will be useful to (re)evaluate these findings. Currently, the full data set is split only once: one training set and one testing set. Future research should look at other means of cross validation that this paper does not explore—For example, $k$-fold cross validation, which generates $k$ random training and testing data sets, thus. maximizing the small data sets available for training and testing by producing more prediction outputs for extreme-high values. Finally, further research might investigate how changing the kernel function influences the effectiveness of the proposed methods. It should also consider which of the two methods, bootstrapping and data expansion, is better at hyperparameter optimization on every training set and compare their effectiveness in SVR and other regression techniques.

In practice, individuals tasked with long-term monitoring of water quality should be aware that the rarity of extreme-high values hinders the development of regression models that accurately predict values for both extreme-high and nonextreme values. They should therefore collect additional data (if possible) during periods when the extreme-high values occur. Ideally, these additional data will be collected at different times when the values of other variables used to build the model are different. However, replicates collected during the same period will be useful in building a prediction model or even serving as backups if an accurate reading cannot be made for the primary sample.

## Conclusion

This study sought to mitigate the systematic underprediction of SVR models used for filling in missing extreme-high BODs by introducing bootstrapping and data expansion methods to increase the proportion of extreme-high values in the training set. For the water quality data set of Yuen Long Creek, Hong Kong, it is found that a systematic underprediction is present for extreme-high BODs of $>150$ mg L$^{-1}$, where such extreme-high values and their SVs are sparsely scattered over most (~70%) of the BOD range in the full data set. The two methods are successful in mitigating systematic underprediction in extreme-high values and in reducing residual errors by up to ~25% as the proportion of extreme-high values in the training set increases to ~40%. These improvements come with a slight worsening of overall model performance, but this negative aspect is tolerable because the extreme-high values are often the missing ones in typical hydrological data sets.

## Appendix. BOD Time Series

This appendix contains BODs for the four water quality monitoring points (YL1, YL2, YL3, and YL4) recorded over the period 2000–2014. They are presented as a time series plot in Fig. 5.

## References

Balfer, J., and Bajorath, J. (2015). "Systematic artifacts in support vector regression-based compound potency prediction revealed by statistical and activity landscape analysis." *PLoS One*, 10(3), e0119301.

Chiang, C. F., Wu, Y. S., and Young, J. C. (2004). "Analyzing the uncorrected error of dilution water demand for the dilution biochemical oxygen demand method." *Water Environ. Res.*, 76(3), 238–244.

Džeroski, S., Demšar, D., and Grbović, J. (2000). "Predicting chemical parameters of river water quality from bioindicator data." *Appl. Intell.*, 13(1), 7–17.

EPDHK (Environmental Protection Department of Hong Kong). (2007). "Livestock waste information system." ⟨http://www.epd.gov.hk/epd/misc/river_quality/1986-2005/eng/5_nor_nt_-menu.htm⟩ (Jan. 21, 2017).

EPDHK (Environmental Protection Department of Hong Kong). (2014). "River water quality in Hong Kong in 2014." ⟨http://wqrc.epd.gov.hk/pdf/water-quality/annual-report/RiverReport2014eng.pdf⟩ (Jan. 21, 2017).

Garsole, P., and Rajurkar, M. (2015). "Streamflow forecasting by using support vector regression." *Proc., 20th Int. Conf. of Hydraulics, Water Resources and River Engineering*, Indian Society for Hydraulics, Pune, India.

Granata, F., Gargano, R., and de Marinis, G. (2016). "Support vector regression for rainfall-runoff modeling in urban drainage: A comparison with the EPA's storm water management model." *Water*, 8(3), 69.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*, Springer, New York.

Hsu, C.-W., Chang, C.-C., and Lin, C.-J. (2003). "A practical guide to support vector classification." ⟨http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf⟩ (Jan. 212017).

Karush, W. (1939). "Minima of functions of several variables with inequalities as side constraints." M.S. thesis, Dept. of Mathematics, Univ. of Chicago, Chicago.

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection." *Proc., Int. Joint Conf. of Artificial Intelligence*, Morgan Kaufmann Publishers, San Francisco, 1137–1145.

Kuhn, H. W., and Tucker, A. W. (2014). "Nonlinear programming." *Traces and emergence of nonlinear programming*, Springer, New York, 247–258.

Lima, A. R., Cannon, A. J., and Hsieh, W. W. (2015). "Nonlinear regression in environmental sciences using extreme learning machines: A comparative evaluation." *Environ. Modell. Software*, 73, 175–188.

Liu, M., and Lu, J. (2014). "Support vector machine—An alternative to artificial neuron network for water quality forecasting in an agricultural nonpoint source polluted river?" *Environ. Sci. Pollut. Res.*, 21(18), 11036–11053.

Nagel, B., Dellweg, H., and Gierasch, L. M. (1992). "Glossary for chemists of terms used in biotechnology (IUPAC recommendations 1992)." *Pure Appl. Chem.*, 64(1), 143–168.

Nash, J., and Sutcliffe, J. (1970). "River flow forecasting through conceptual models. Part I: A discussion of principles." *J. Hydrol.*, 10(3), 282–290.

Noori, R., Karbassi, A., Ashrafi, K., Ardestani, M., Mehrdadi, N., and Bidhendi, G.-R. N. (2012). "Active and online prediction of BOD5 in river systems using reduced-order support vector machine." *Environ. Earth Sci.*, 67(1), 141–149.

Noori, R., Yeh, H.-D., Abbasi, M., Kachoosangi, F. T., and Moazami, S. (2015). "Uncertainty analysis of support vector machine for online prediction of five-day biochemical oxygen demand." *J. Hydrol.*, 527, 833–843.

Qiu, J.-W. (1999). "Composition, structure and distribution of polychaete assemblages in Deep Bay." *The mangrove ecosystem of deep bay and the Mai Po marshes, Hong Kong*, Hong Kong University Press, Hong Kong, 13–21.

Rice, E., Baird, R., Eaton, A., and Clesceri, L. S. (2012). *Standard methods for the examination of water and wastewater*, American Public Health Association, American Water Works Association, Water Environment Federation, Washington, DC.

Sawyer, C. N., McCarty, P. L., and Parkin, G. F. (2002). *Chemistry for environmental engineering and science*, 5th Ed., McGraw Hill, New York.

Singh, K. P., Basant, A., Malik, A., and Jain, G. (2009). "Artificial neural network modeling of the river water quality—A case study." *Ecol. Modell.*, 220(6), 888–895.

Smola, A. J., and Scholkopf, B. (2004). "A tutorial on support vector regression." *Stat. Comput.*, 14(3), 199–222.

Udeigwe, T. K., and Wang, J. J. (2010). "Biochemical oxygen demand relationships in typical agricultural effluents." *Water Air Soil Pollut.*, 213(1–4), 237–249.

Vapnik, V. N. (1995). "Constructing learning algorithms." *The nature of statistical learning theory*, Springer, New York, 119–166.