Environmental Pollution 317 (2023) 120734

Contents lists available at ScienceDirect

Environmental Pollution

journal homepage: www.elsevier.com/locate/envpol

# Using machine learning models to predict the effects of seasonal fluxes on *Plesiomonas shigelloides* population density

Temitope C. Ekundayo [a,b,c,*], Oluwatosin A. Ijabadeniyi [b], Etinosa O. Igbinosa [a,d], Anthony I. Okoh [a,e]

[a] SAMRC Microbial Water Quality Monitoring Centre, University of Fort Hare, Alice, Eastern Cape, South Africa
[b] Department of Biotechnology and Food Science, Durban University of Technology, Steve Biko Campus, Steve Biko Rd, Musgrave, Berea, 4001, Durban, South Africa
[c] Department of Microbiology, University of Medical Sciences, Ondo City, Ondo State, Nigeria
[d] Department of Microbiology, Faculty of Life Sciences University of Benin, Private Mail Bag 1154, Benin City, 300283, Nigeria
[e] Department of Environmental Health Sciences, College of Health Sciences, University of Sharjah, Sharjah, P.O. Box 27272, United Arab Emirates

ARTICLE INFO

*Keywords:*
Pathogen
Public health
Machine intelligence
Prediction
Feature importance
Predictive microbiology
Multiple linear regression
Random forest
Gradient boosted machine
Neural networks
K-nearest neighbours
Boosted regression tree
Extreme gradient boosted regression
Support vector regression
Decision tree regression
M5 pruned regression
Artificial neural network regression

ABSTRACT

Seasonal variations (SVs) affect the population density (PD), fate, and fitness of pathogens in environmental water resources and the public health impacts. Therefore, this study is aimed at applying machine learning intelligence (MLI) to predict the impacts of SVs on *P. shigelloides* population density (PDP) in the aquatic milieu. Physicochemical events (PEs) and PDP from three rivers acquired via standard microbiological and instrumental techniques across seasons were fitted to MLI algorithms (linear regression (LR), multiple linear regression (MR), random forest (RF), gradient boosted machine (GBM), neural network (NN), K-nearest neighbour (KNN), boosted regression tree (BRT), extreme gradient boosting (XGB) regression, support vector regression (SVR), decision tree regression (DTR), M5 pruned regression (M5P), artificial neural network (ANN) regression (with one 10-node hidden layer (ANN10), two 6- and 4-node hidden layers (ANN64), and two 5- and 5-node hidden layers (ANN55)), and elastic net regression (ENR)) to assess the implications of the SVs of PEs on aquatic PDP. The results showed that SVs significantly influenced PDP and PEs in the water ($p < 0.0001$), exhibiting a site-specific pattern. While MLI algorithms predicted PDP with differing absolute flux magnitudes for the contributing variables, DTR predicted the highest PDP value of 1.707 log unit, followed by XGB (1.637 log unit), but XGB (mean-squared-error (MSE) = 0.0025; root-mean-squared-error (RMSE) = 0.0501; $R^2$ =0.998; medium absolute deviation (MAD) = 0.0275) outperformed other models in terms of regression metrics. Temperature and total suspended solids (TSS) ranked first and second as significant factors in predicting PDP in 53.3% (8/15) and 40% (6/15), respectively, of the models, based on the RMSE loss after permutations. Additionally, season ranked third among the 7 models, and turbidity (TBS) ranked fourth at 26.7% (4/15), as the primary significant factor for predicting PDP in the aquatic milieu. The results of this investigation demonstrated that MLI predictive modelling techniques can promisingly be exploited to complement the repetitive laboratory-based monitoring of PDP and other pathogens, especially in low-resource settings, in response to seasonal fluxes and can provide insights into the potential public health risks of emerging pathogens and TSS pollution (e.g., nanoparticles and micro- and nanoplastics) in the aquatic milieu. The model outputs provide low-cost and effective early warning information to assist watershed managers and fish farmers in making appropriate decisions about water resource protection, aquaculture management, and sustainable public health protection.

## 1. Introduction

Parallel (re)emergence of freshwater pathogens is one of the major global concerns of millennials (Ostfeld et al., 2010). *P. shigelloides* is particularly ignored among the emerging pathogens. The implications of this pathogen in several veterinary and human clinical cases, such as central nervous system infections and gastroenteritis–one of the contributing causes of infant death globally (WHO, 2009; CDC, 2015)–call for critical attention and wide surveillance studies of the pathogen. An increase in *Plesiomonas* gastroenteritis has been reported, with a high

* Corresponding author. SAMRC Microbial Water Quality Monitoring Centre, University of Fort Hare, Alice, Eastern Cape, South Africa.
*E-mail address:* cyruscyrusthem@gmail.com (T.C. Ekundayo).

https://doi.org/10.1016/j.envpol.2022.120734
Received 7 June 2022; Received in revised form 21 November 2022; Accepted 22 November 2022
Available online 28 November 2022

prevalence in Africa and Southeast Asia (Chen et al., 2013). *P. shigelloides* not only causes human infection but also has been recognized as one of the major pathogens associated with clinical cases of cultured fish, including cultured sturgeons (Wang et al., 2013), *Ctenopharyngodo nigellas* (Hu et al., 2014), *Oreochromis niloticus* (Liu et al., 2015; Sierralta Chichizola et al., 2016), *Hypophthalmichthys molitrix* (Behera et al., 2018), *Acipenser dabryanus* (Jiang et al., 2021), and *Cyprinus carpio* (Chen et al., 2022), resulting in devastating mortality.

*P. shigelloides* is faecal-orally transmitted to humans via ingestion of contaminated products such as vegetables, water, and seafood/aquatic foods and occupational contact with amphibians and reptiles in a freshwater milieu (Shinohara et al., 2021; Janda et al., 2016; Ciznar et al., 2006; Keating, 2005). Increasing reports have shown that *P. shigelloides* possesses virulence traits similar to those of other pathogens, such as cholera-like toxins (Gardner et al., 1987), enterotoxins (Abbott et al., 1991), β-heamolysin (Janda and Abbott, 1993), and the cytotoxin lipopolysaccharide complex (Okawa et al., 2004). The pathogen has implications in infections in individuals of certain extreme ages (especially infants and the elderly) (Stock, 2004) and immunocompromised health status (Stock, 2004; Obi et al., 2002), in poor sanitary conditions (Jagger, 2000), and in areas with high aquatic temperature (Stock, 2004). Although its pathogenicity has not been confirmed from a Koch's postulates perspective, horizontal gene transfer has been linked to the increasing pathogenicity of *P. shigelloides* in a recent study (Yin et al., 2020), and that ignoring this information could lead to significant clinical and public health concerns.

*P. shigelloides* has been isolated in aquatic systems (Janda et al., 2016). However, there is scarce information on the purposeful investigation of the effects of seasonal variations (SVs) on *P. shigelloides* population density (PDP) in freshwaters. Few studies, however, reported an increase in *Plesiomonas* infections during the summer period in the tropics (Janda et al., 2016). Since the environmental fate, spread, and persistence of microbial pathogens are often linked with SVs and physicochemical fluxes (Vezzulli et al., 2013; Bonadonna et al., 2002), the relationship between pathogen population density (PD) and SVs in physicochemical events (PEs) is a significant factor in the epidemiology of their infections. Knowledge of the SVs in PEs in freshwater could facilitate the forecasting and prevention of waterborne disease outbreaks (Vezzulli et al., 2013; Gutiérrez-Salazar et al., 2011). The significant direct and indirect impacts of meteorological phenomena (extreme precipitation, flooding, environmental deterioration, extreme temperature and environmental cycles) on aquatic PEs have also been demonstrated (Tong and Lu, 2000).

In addition, assessing PDP using *in situ* and laboratory measurements is time-consuming, labour intensive and cost intensive. Therefore, the establishment of a predictive framework and intelligent system for determining PDP may indicate to be a promising, low-cost, effective, early warning alternative system to assist fish farmers, watershed managers, and public health protection against the pathogen. The use of machine learning (ML) algorithms in developing predictive decision systems for assessing microbial pathogens and ecological and anthropological impacts is gaining popularity, and such methods have been widely applied. For instance, DeLuca et al. (2020) predicted the presence/absence and abundance of *Vibrio parahaemolyticus* in marine environments using ML, such as with a support vector machine/support vector regression (SVM/SVR) and a random forest (RF). RFs have also been applied in determining metal resistance genes in estuary ecosystems and estuarine bacteria and in predicting environmental health, functional gene levels, and antibiotic resistance genes based on 16 S rRNA sequencing, microbiome, and metagenomics data (Sun et al., 2021; Wilhelm et al., 2022; Zhou et al., 2022). RFs were employed in microbial host source tracking of nonstochastic and divergent assembly of gut microbiomes in farmed and wild large yellow croaker (Zhu et al., 2022). In another study, RFs were used to predict the most influential factors that had affinity with *Salmonella* detection in agricultural watersheds (Toro et al., 2022). Other ML methods that have been utilized to predict microbial-related events include M5 pruned regression (M5P) and extreme learning machines (ELMs) for algal bloom prediction (Yi et al., 2019); artificial neural networks (ANNs); SVMs (Park et al., 2015; Lee et al., 2003); K-nearest neighbours (kNN) (Ye et al., 2014); neural network-based MLI for predicting necrotizing enterocolitis from premature infant faecal microbiota (Lin et al., 2022); KNN, RFs, SVMs, and deep neural networks (DNNs) in assessing quorum-sensing among human gut microbiota (Wu et al., 2022); and soil fungal diversity and abundance using RFs, Cubist, SVMs, extreme gradient boosting (XGB), Gaussian process regression, and one-dimensional convolutional neural networks (Yang et al., 2022). Additionally, MLI, such as SVR, ELM, classification and regression tree (CART), RF, ANNs, linear regression (LR), and KNN algorithms, have been applied in building water management facilities and smart control systems (Amini et al., 2021). ML has been applied for groundwater nitrate contamination level (Rahmati et al., 2019; Band et al., 2020) and groundwater level (Kombo et al., 2020; Chen et al., 2020) estimations.

To the best of our knowledge, no information exists about the impacts of SVs and resultant PEs on PDP in the aquatic milieu or its investigation in South Africa, as well as its intelligent-based model for decision-making. Therefore, the present investigation was aimed at assessing MLI in modelling the implications of SVs of PEs on PDP in an aquatic milieu in Eastern Cape Province, South Africa.

## 2. Materials and methods

### 2.1. River sites and sample collection

The chosen sampling sites were situated on the Kat River, Tyhume River, and Kubusi River in the Raymond Mhlaba Local Municipality, as previously described by Ekundayo and Okoh (2019). The selected rivers were sampled across seasons. Seasons were defined as late summer to mid-autumn (February–April), late autumn to mid-winter (May–July), late winter to mid-spring (August–October), and late spring to early summer (November–December) to reflect the actual sampling period in this study. However, the normal seasons in the South Africa period are December–February (summer), March–May (autumn), June–August (winter), and September–November (spring). Five essential sampling locations were selected from each river. In all, three replicates of midstream water (25–30 cm depth) were sampled into sterile glass bottles from every sampling site and transported in iceboxes to the laboratory for further analysis. All the samples throughout the entire study period were analysed within 6 h of collection (American Public Health Association, 2005).

### 2.2. Data acquisition

The assessment of PEs and PDP was as described by Ekundayo and Okoh (2019). The PEs that were measured included pH, temperature (°C), total suspended solids (TSS, mg/L), DO (dissolved oxygen, mg/L), TDS (total dissolved solids, mg/L), TBS (turbidity, nephelometric turbidity unit (NTU)), electrical conductivity (EC, μs/cm), salinity (practical salinity unit, PSU) and biochemical oxygen demand (BOD). The PDP in the water samples was estimated by using a 0.45 μm pore-sized filter (Ø 47 mm) for membrane filtration of 100 ml sample diluents plated onto inositol brilliant green bile agar at 39 °C for 24 h for the characteristic pink colonies (X) of *P. shigelloides*. Additionally, *P. shigelloides* colonies were validated by a randomized oxidase test and *Plesiomonas*-specific-polymerase chain reaction (PCR).

### 2.3. Model development

#### 2.3.1. Preprocessing

The PDP data were log-transformed. All PE datasets were centre-scaled (mean = 0 and standard deviation = 1) in the models. The four prevailing seasons in South Africa were coded as 1–4 (1 = late summer

to mid-autumn (LSMA), 2 = late autumn to mid-winter (LAMW), 3 = late winter to mid-spring (LWMS), and 4 = late spring to early summer (LSES)) as a reflection of the actual sampling period in this study.

#### 2.3.2. Analysis and modelling procedure

First, explanatory analysis was performed on data using beanplots comparative visualization of the site-specific variable distribution in R v.3.4.4 software (2017-06-30). The beanplots allow the representation of individual observations in a dataset as small lines on a 1-dimensional scatter plot and can be utilized to estimate the distributions' density and to determine whether a group contains adequate observations for statistical implementation. Additionally, duplicate measurements, bimodal distributions and other anomalies in data are easily recognized in beanplots (Kampstra, 2008). The study compared seasonal PEs and PDP with a composite violin and box plot using ggpubr r package version 0.1.7 (Kassambara, 2018; https://CRAN.R-project.org/package=ggpubr).

Second, the study modelled the SVs of PDP as a dependent variable of the seasonal distribution of PEs. The conditional expected value of PDP given PEs, the vector of values of PEs (pH, conductivity, TDS, salinity, temperature, TSS, DO, and BOD) for a particular season (z), is expressed as $\mathbf{E}_{\boldsymbol{PDP}|\boldsymbol{PEs}|z}(\boldsymbol{PDP})$. Therefore, the approximation of the PDPs (mean) value can be modelled as $\mathbf{E}_{(\boldsymbol{PDPPEs}|z)(\boldsymbol{PDP})} \approx \boldsymbol{f}(\boldsymbol{PEs}|z)$.

Based on the foregoing, 15 regression-based models (LR, multiple linear regression (MR), an RF, a gradient boosted machine (GBM), neural network (NNT) (10-6-1 network with 73 wt), a KNN, a boosted regression tree (BRT), XGB, SVR, a decision tree regression (DTR), M5P, ANN regression (with one 10-node hidden layers (ANN10), two 6- and 4-node hidden layers (ANN64), and two 5- and 5-node hidden layers (ANN55)), and elastic net regression (ENR)) were fitted to the acquired data (511 observations with 10 variables) by inputting 70% of the data as a learning dataset for estimation of the model coefficients and the remaining 30% as a validation dataset for model validation. The models were selected for their ability to fit multidimensional variables for regression tasks on both ordinal and continuous response variables. All models were implemented with 10-fold cross-validation coupled with 3 repeats and a tune-length set to 10, generating ten different learning and validation dataset pairs. In all cases, optimal hyperparameters were identified and chosen via a grid search. The detailed model parameters are presented in the supplemental material.

In all ML model cases, a breakdown (BD) plot was utilized as the explanatory output for the model, as explained by Equation (1):

$$f(x.) = v_0 + \sum_{j=1}^{p} v(j, x.) \tag{1}$$

where $v(j, x.)$ denotes the jth variable contribution or variable-importance measure to the model's prediction at instance $x$ and $v_0$ is the mean model prediction (refer to Biecek and Burzykowski, 2021, for full derivation).

### 2.4. Evaluation of model performance and sensitivity analysis

The model-predicted PDP values by the MLI algorithms were examined against experimental data using the mean squared-error (MSE), root-mean-square error (RMSE), $R^2$, median absolute deviation (MAD), and the RMSE was also applied as a loss function for training and confirmation errors (Namkung, 2020; HANSEN, 1993).

$$\boldsymbol{MSE}\left(\boldsymbol{f}, \underline{\boldsymbol{X}}, \underline{\boldsymbol{y}}\right) = \frac{\mathbf{1}}{\boldsymbol{k}} \sum_{i}^{k} (\hat{y}_i - y_i)^2 = \frac{\mathbf{1}}{\boldsymbol{k}} \sum_{1}^{k} \boldsymbol{r}_i^2 \tag{2}$$

$$\boldsymbol{RMSE}\left(\boldsymbol{f}, \underline{\boldsymbol{X}}, \underline{\boldsymbol{y}}\right) = \sqrt{\boldsymbol{MSE}\left(\boldsymbol{f}, \underline{\boldsymbol{X}}, \underline{\boldsymbol{y}}\right)} \tag{3}$$

$$\boldsymbol{R}^2\left(\boldsymbol{f}, \underline{\boldsymbol{X}}, \underline{\boldsymbol{y}}\right) = \mathbf{1} - \frac{\boldsymbol{MSE}\left(\boldsymbol{f}, \underline{\boldsymbol{X}}, \underline{\boldsymbol{y}}\right)}{\boldsymbol{MSE}\left(\boldsymbol{f}_{\mathbf{0}}, \underline{\boldsymbol{X}}, \underline{\boldsymbol{y}}\right)} \tag{4}$$

$$\boldsymbol{MAD}\left(\boldsymbol{f}, \underline{\boldsymbol{X}}, \underline{\boldsymbol{y}}\right) = \boldsymbol{median}(|\boldsymbol{r}_1|, \ldots, |\boldsymbol{r}_n|) \tag{5}$$

where $k$ = number of the sample; $f_0()$: baseline model; $r_i$: residual for the $i_{th}$ observation, $\underline{X}$: matrix of explanatory variables; $\underline{y}$: vector of the values of the dependent variable; $\boldsymbol{f}(\underline{\hat{\boldsymbol{\theta}}}, \underline{\boldsymbol{X}})$: model based on the training dataset; $\underline{\hat{\boldsymbol{\theta}}}$: estimated values of the model's coefficients; and $\underline{\hat{y}}_i$: model's prediction corresponding to $y_i$.

The sensitivity of the models was determined via residual diagnostics (i.e., comparing input data with the predicted value) and partial-dependence profiles (P) of the PDP on the predictor variables. The partial-dependence profile of a Model $f()$ (i.e., expected predicted PDP value by the model) and the predictive variable $X^j$ fixed at $m$ (over the marginal/empirical distribution of $X^j$ (n), i.e., over the joint distribution of all predictive variables other than $X^j$) is estimated by Equations (6) and (7):

$$\boldsymbol{h}_{\boldsymbol{P}}^{\boldsymbol{j}}(\mathbf{m}) = \boldsymbol{E}_{\underline{\boldsymbol{X}}^{-j}}\left\{\boldsymbol{f}\left(\boldsymbol{X}^{j|=\boldsymbol{m}}\right)\right\} \tag{6}$$

$$\widehat{\boldsymbol{h}}_{\boldsymbol{P}}^{\boldsymbol{j}}(\mathbf{m}) = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{i=1}^{n} \boldsymbol{f}\left(\underline{\boldsymbol{x}}_{\boldsymbol{i}}^{j|=\boldsymbol{m}}\right) \tag{7}$$

All models were implemented in R v.4.1.2 software.

## 3. Results

### 3.1. Location-specific variation/distribution of PEs and PDPs

The distributions of PEs and PDP varied across sampling locations throughout the four seasons (Figure S1). All parameters were significantly different across the locations ($p < 0.05$). The pH of the freshwater significantly varied ($p < 0.05$). While the lowest pH on the Tyhume River was 7.25 (at Binfield), the highest pH was 7.86 (Hogsberg) (Figure S1(a)). At Fortbeaufort, the highest pH (7.48) was measured on the Kat River, and pH 7.71 (highest pH) was measured at StuttFbridg on the Kubusie River. The average TBS across the sampling sites ranged from 12.85 NTU (Binfield) to 91.45 NTU (StuttTW1). The highest mean value of TBS in the Kat River, Tyhume River, and Kubusie River was 42.55 NTU (Seymour), 29.15 NTU (Melani) and 91.45 NTU (StuttTW1), respectively (Figure S1(b)).

The mean value of TDS varies from 40.24 mg/l (Hogsberg) to 261 mg/l (StuttGResrv). The EC had the lowest value at Katberg (135 μs/cm) on the Kat River, at Hogsberg (81 μs/cm) on the Tyhume River, and at StuttEbBridg (196 μs/cm) on the Kubusie River (Figure S1(c)). However, the EC was highest at Seymour (362 μs/cm), Melani (159 μs/cm) and StuttGResrv (522 μs/cm) on the Kat River, Tyhume River, and Kubusie River, respectively.

Additionally, the average TSS varied from 10.27 mg/l (Binfield) to 82 mg/l (StuttTW1), and the mean BOD ranged from 1.34 mg/l (Hogsberg) to 3.38 mg/l (Kayaletu) throughout the locations (Figure S1(d)). The mean salinity was high at StuttFbridg (1.30 PSU) compared with other locations (Figure S1(e)). The mean seasonal temperature of the freshwater ranged from 11.14 °C (Hogsberg) to 19 °C (Binfield) ($P < 0.05$). The highest mean temperature was measured at Fortbeaufort (17 °C), Binfield (19 °C) and StuttEbBridg (17 °C) on the Kat River, Tyhume River and Kubusie River, respectively (Figure S1 (g)). Similarly, the mean DO varied between 6.69 mg/l (Seymour) and 9.58 mg/l (Hillfoot) throughout the sampling sites (Figure S1(h)).

The average PDP varied from 0.71 log units (Hillfoot) to 1.89 log units (StuttFbridg) (Figure S1(g)). Generally, a high PDP was associated with some sites, including Blinkwater (1.40 log unit) and Fortbeaufort

(1.32 log unit) on the Kat River; Binfield (1.61 log unit) and Melani (1.76 log unit) on the Tyhume River; and StuttEbBridg (1.87 log unit) and StuttFbridg (1.89 log unit) on the Kubusie River (see Fig. 1).

### 3.2. River-specific seasonal distribution of PEs and PDP

Fig. 2 shows the seasonal distribution of PDP and PEs in freshwater. Seasonal water temperatures were significantly different ($F = 52.454$, $P \leq 0.05$) (Fig. 1a). In LSMA, the water temperature ranged from 12.7 to 25.8 °C, 4.7–19 °C, and 5.3–20.7 °C in the Tyhume River, Kubusie River, and Kat River, respectively. The average water temperature in LAMW was generally low in the Tyhume River (16.5 ± 3.2 °C, B1), Kubusie River (18.4 ± 1.8 °C, B2) and Kat River (10.7 ± 2.7 °C: B3) compared with other periods. Similarly, relatively low temperatures characterized LWMS in the Tyhume River (6.5–12.6 °C, C1), Kubusie River (15.9–21.0 °C, C2), and Kat River (14.9–23.6 °C, C3). However, high temperatures were recorded in LSES from 8.0 to 16.6 °C (D1), 12.3–18.8 °C (D2), and 17.8–24.0 °C (D3) in the Tyhume River, Kubusie River, and Kat River, respectively.

Fig. 1b shows the significantly different seasonal values of DO (mg/l) in the freshwater resources ($F = 19.368$, $P \leq 0.05$). DO levels were generally high in LAMW and LWMS compared with another period. DO in the Tyhume River, Kubusie River, and Kat River ranged in order from 7.66 to 9.20 mg/l (A1), 8.75–10.61 mg/l, and 7.02–11.11 mg/l (A3), respectively, in LSMA and from 7.8 to 9.96 mg/l (B1), 2.57–9.09 mg/l (B2), and 1.65–11.38 mg/l (B3), respectively, in LAMW. DO ranged from 6.34 to 12.15 mg/l (C1), 5.28–9.04 mg/l (C2), and 5.69–9.52 mg/l (C3) during LWMS in the Tyhume River, Kubusie River, and Kat River, respectively. During LSES, DO varied from 6.02 to 11.09 mg/l (D1), 7.81–10.24 mg/l (D2), and 4.94–9.66 mg/l (D3) in the Tyhume River, Kubusie River, and Kat River, respectively.

The seasonal pH values of the rivers were significantly different ($F = 20.229$, $P \leq 0.05$) (Fig. 1c). The pH values generally fluctuated throughout the seasons, with an overall mean of 7.49 ± 0.44. High pH values characterized LSES in all the rivers. Specifically, pH varied from 6.8 to 8.1 (A1), 6.8–8.7 (A2), and 6.7–8.8 (A3) during LSMA in the Tyhume River, Kubusie River, and Kat River, respectively, and from 7.4 to 9.4 (B1), 6.3–7.8 (B2), and 6.6–7.6 (B3) during LAMW in the Tyhume River, Kubusie River, and Kat River, respectively. Additionally, during LWMS, pH fluctuated between 6.8 and 7.8 (C1), 7.5 and 9.2 (C2), and 7.1 and 8.4 (C3) in the Tyhume River, Kubusie River, and Kat River, respectively. In LSES in the Tyhume River, Kubusie River, and Kat River, the pH ranged from 7.1 to 7.9 (D1), 7.1–8.1 (D2) and 7.4–7.9 (D3), respectively.

Fig. 1d gives the TDS in the freshwater ($F = 16.059$, $P \leq 0.05$). The mean ± SE values of TDS during LSMA were 51.71 ± 2.55 mg/l (A1, Tyhume River), 58.36 ± 3.12 (A2: Kubusie River), and 55.80 ± 3.63 (A3: Kat River). The TDS values of 56.91 ± 3.25 mg/l (B1), 81.45 ± 5.46 mg/l (B2), and 96.60 ± 5.45 mg/l (B3) were recorded in the Tyhume River, Kubusie River, and Kat River, respectively, during LAMW. The TDS values were lower in the Tyhume River in LWMS (112 ± 8 mg/l (C1)) than in LSES (168 ± 20 (D1)). While the mean TDS was 152 ± 15 (C2) in LWMS, the TDS was 142 ± 18 mg/l (D2) during LSES in the Kubusie River. The Kat River had an average value of 108 ± 10 mg/l (C3) in LWMS and of 170 ± 21 mg/l (D3) in LSES.

The distribution of TBS in the freshwaters was significantly different during the seasons ($F = 5.976$, $P \leq 0.05$) (Fig. 1e). In the Tyhume River, TBS ranged between 5 and 56 mg/l (A1) in LSMA, 6 and 95 mg/l (B1) in LAMW, 3 and 27 mg/l (C1) in LWMS, and 1 and 239 mg/l (D1) in LSES. Likewise, Kubusie River's TBS was 5–33 mg/l (A2), 10–71 mg/l (B2), 0–80 mg/l (C2), and 0–135 mg/l (D2) during LSMA, LAMW, LWMS, and LSES, respectively. In the Kat River, TDS ranged from 4 to 877 mg/l (A3) during LSMA, 1–211 mg/l (B3) during LAMW, 5–296 mg/l (C3) during LWMS, and 0–80 mg/l (D3) during LSES.

Fig. 1f presents the different seasonal EC levels in the rivers ($F = 16.516$, $P \leq 0.05$). In all seasons, the Tyhume River had the lowest EC (mean ± SE) compared with the other rivers. The average EC was 98.31 ± 4.47 μs/cm in LSMA, 114 ± 6 μs/cm (B1) in LAMWA, 224 ± 17 μs/cm (C1) in LWMS, and 336 ± 40 μs/cm (D1) in LSES in the Tyhume River. However, relatively high EC values were recorded in the Kubusie River and Kat River in all seasons. EC was 117 ± 6 μs/cm (A2) in the Kubusie River and 106 ± 8 μs/cm (A3) in the Kat River during LAMW. The EC values during LAMW and LWMS were 163 ± 11 μs/cm (B2) and 304 ± 31 μs/cm (C2), respectively, in the Kubusie River, whereas the EC values in the Kat River were 190 ± 11 μs/cm (B3) in LAMW and 216 ± 19 μs/cm (C3) during LWMS. Other values of EC include 290 ± 34 μs/cm (D2) and 314 ± 37 μs/cm (D3).

The TSS distribution was significantly different ($F = 6.017$, $P \leq 0.05$) in the rivers during the survey (Fig. 1g). TSS ranged from 3.0 to 51 mg/l (A1), 1–51 mg/l (A2), and 1–829 mg/l (A3) in LSMA; 5–85 mg/l (B1), 6–68 mg/l (B2), and 1–185 mg/l (B3) during LAMWA; 3–22 mg/l (C1), 0–70 mg/l (C2), and 4–286 mg/l (C3) in LWMS; and 1–217 mg/l (D1), 0–125 mg/l (D2), and 0–70 mg/l (D3) during LSES.

The BOD of the river samples varied according to the season ($F = 9.436$, $P \leq 0.05$; Fig. 1h). BOD varied from 0.9 to 4.69 mg/l, 0.15–4.35 mg/l, and 0.95–9.04 mg/l at A1, A2 and A3, respectively. Similarly, the BOD ranges were 0.71–8.11 mg/l (B1), 1.05–7.04 mg/l (B2), 0.39–3.68 mg/l (B3), 1.02–4.51 mg/l (C1), 2.22–5.41 mg/l (C2), 0.81–5.75 mg/l (C3), 0.54–5.47 mg/l (D1), 0.92–7.52 mg/l (D2), and 1.17–6.79 mg/l (D3) for the respective seasons and locations. The salinity varied across the seasons in the freshwater ($F = 15.678$, $P \leq 0.05$; Fig. 1i). Mostly, the values vary from 0.02 to 0.07 PSU (A1), 0.03–0.12 PSU (A2), 0.02–0.1 PSU (A3), 0.03–0.08 PSU (B1), 0.04–0.18 PSU (B2), 0.05–0.19 PSU (B3), 0.06–0.29 PSU (C1), 0.05–0.35 PSU (C2), 0.04–0.24 PSU (C3), 0.04–0.41 PSU (D1), 0.04–0.37 PSU (D2), and 0.05–0.35 PSU (D3) at the respective locations and seasons.

Highly variable PDP was observed throughout the **seasons** ($F = 36.695$, $P \leq 0.05$; Fig. 1j), with a greater number of values associated with LSMA and LSES than with LAMWA and LWMS. The PDP values included 0–3.04 log units (A1), 0–1.61 log units (A2), 0–2.59 log units (A3), 2–3.23 log units (B1), 0–3.04 log units (B2), 0–2.3 log units (B3), 0–2.28 log units (C1), 0.7–3.23 log units (C2), 0–4.15 log units (C3), 0–2.48 log units (D1), 0–3.08 log units (D2), and 0.7–3.45 log units (D3).

### 3.3. Model predicted PDP values and descriptive contribution of variables under seasonal/PE fluxes in the aquatic milieu

The model-predicted PDP by the MLI models is shown in Table 1. The predicted PDP ranged from 0.170 to 1.707 log units. DTR predicted the highest PDP (1.707 log unit), followed by XGB (1.637 log unit), the GBM (1.618 log unit), the BRT (1.397 log unit), the RF (1.326 log unit), M5P (1.225 log unit), SVR (1.217 log unit), and MR (1.126 log unit), among others.

Fig. 2 presents an explanatory plot of the contributions of the various variables to the models' prediction of PDP in the aquatic milieu. The absolute contributions of the variables differ by model. While LR identified TSS (0.38), salinity (0.371) and TBS (0.32) as variables with a greater contribution, MR indicated salinity (0.604), conductivity (0.462) and TSS (0.444) as the best contributing PEs in decreasing order of flux magnitudes. KNN identified pH (0.168), BOD (0.145), and temperature (0.116) as having higher contributions to predicting PDP in descending order. The contributions of PEs in the RF's prediction of PDP increase in flux magnitude from temperature (0.174), DO (0.219), and pH (0.258) as the best 3 variables. The first two PEs with the highest contributions to PDP prediction and corresponding flux magnitude by XGB (DO (0.394), BOD (0.356)), BRT (temperature (0.285), TSS (0.236)), GBM (temperature (0.298), DO (0.111)), NNT (TDS (0.987), conductivity (0.57)), DTR (DO (0.505), TSS (0.145)), SVR (temperature (0.194), salinity (0.168)), M5P (temperature (0.278), DO (0.239)), ENR (salinity (0.368), conductivity (0.283)), ANN64 (TDS (0.343), temperature (0.242), ANN55 (pH (0.135), season (0.127)), and ANN10 (season (0.155), TSS (0.111)) are presented in parentheses following the

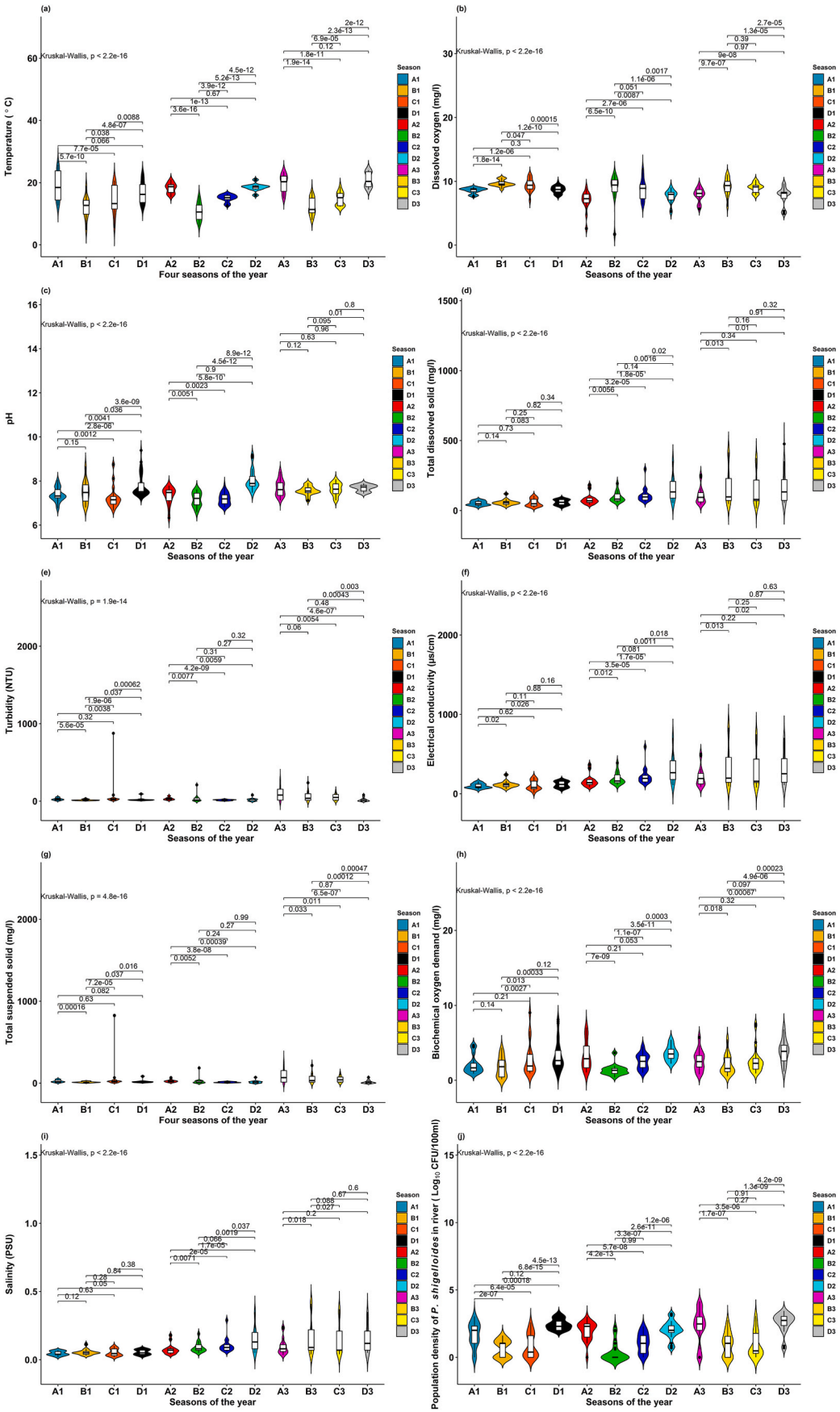


**Fig. 1.** Seasonal distribution of physicochemical variables and PDP in **freshwater**. Seasons are defined as A1, A2 and A3 represented LSMA (February–April) for the Tyhume River, Kubusie River, and Kat River respectively; B1, B2, and B3 represented LAMW (May–July) for the Tyhume River, Kubusie River and Kat River respectively; C1,C2, and C3 represented LWMS (August–October) for the Tyhume River, Kubusie River and Kat River respectively; and D1, D2 and D3 represented LSES (November–December) for the Tyhume River, Kubusie River, and Kat River respectively. The normal South African season is given as summer (December, January, and February), autumn (March, April, and May), winter (June, July, and August), and spring (September, October, and November). In the violin boxplot composite, the violin area presents a distribution including outliers (mild and extreme). In each inner box plot, the central point = median value, and the rectangle = interval between the 25% percentile and 75% percentile. The p-values for the seasonal variables' comparison are annotated on the lines above the bins.

**Table 1**
Model-predicted PD of *P. shigelloides* in the aquatic milieu.

| Rank | Models | PDP (log unit) |
|---|---|---|
| 1 | DTR | 1.707 |
| 2 | XGB | 1.637 |
| 3 | GBM | 1.618 |
| 4 | BRT | 1.397 |
| 5 | RF | 1.326 |
| 6 | M5P | 1.225 |
| 7 | SVR | 1.217 |
| 8 | MR | 1.126 |
| 9 | LR | 1.081 |
| 10 | ENR | 1.058 |
| 11 | KNN | 0.945 |
| 12 | NNT | 0.606 |
| 13 | ANN10 | 0.286 |
| 14 | ANN64 | 0.200 |
| 15 | ANN55 | 0.170 |

respective model.

### 3.4. Model performance of ML algorithms in predicting PDP under seasonal and PE fluxes

The MLI had varied performance in forecasting PDPs under SVs and physicochemical fluxes in aquatic environments (Table 2). For all predictive model-performance indices, XGB (MSE = 0.0025, RMSE = 0.0501; $R^2 = 0.998$, MAD = 0.0275) outperformed the other models in assessing PDPs. In terms of MSE and RMSE, ANN64, ANN10, ANN55, RF, and M5P recorded good performance. In terms of the R-squared measure, RF ($R^2 = 0.9725$), M5P ($R^2 = 0.9717$), ANN64 ($R^2 = 0.9147$), and ANN10 ($R^2 = 0.9098$) had good performance among the models.

### 3.5. MLI model variable importance, feature selection and interpretation

Table 3 (Figure S2) presents the mean variable importance of the seasonal PE variations over 100 permutations across the MLIs in predicting PDP in the aquatic milieu. Based on the RMSE loss over the 100 permutations (Table 3 and Table S), temperature was ranked the most significant PE in predicting PDP in 53.3% (8/15) of the models (i.e., XGB, RF, BRT, KNN, GBM, DTR, SVR, and M5P). Additionally, TSS ranked the most important variable in 6 (40%) models (LR, MR, ENR, ANN55, ANN64, and ANN10). Season ranked second in 7 (XGB, RF, BRT, KNN, DTR, SVR, and M5P) of 15 models, whereas TBS ranked second in 26.7% (4/15: LR, MR, ANN64, and ANN10) of the models. TDS and conductivity ranked first and second, respectively, in NNT as significant variables in predicting PDP in the aquatic milieu.

The resulting plot from the comparison of the observed and predicted PDP values (residual diagnostics) by all models is shown in Fig. 3. This result indicated that the predicted values were closest to the observed values of PDP for XGB (E) AND RF (D). The results also showed that the smoothed trends for M5P (K), ANN10 (O), ANN64 (N), and ANN55 (M) resembled a straight line with a gradient smaller than 1. Generally, the smoothed trends in the models were above and below the straight line at higher values and lower values, respectively, but approximately overlapped in XGB across all values of observed PDP.

Fig. 4 presents the contractive partial-dependence profiles for the fifteen models on PDP and predictive variables. A discretized presentation of Fig. 4 subpanels is given in Figure S3 for clarity. Generally, the shape or trend of the partial-dependence profiles of the models exist in 4 forms: (i) upwards trend from left to right (average prediction increases as the predictor variable increases), (ii) downwards trend from left to right (average prediction decreases as the predictor variable increases), (iii) horizontal trend from left to right (no change in average prediction as the predictor variable increases or decreases), and (iv) a combination of two or three trends.

The shape of the partial-dependence profiles for pH was similar for KNN, RF, XGB, BRT, M5P, GBM, and ANN64 and implies a nonlinear increasing predicted PDP value for an increasing pH until a threshold was reached (7.5 or 8.0), after which a decreasing relationship was observed (Fig. 4 (A) and Figure S3). SVR and ANN10 showed a nonlinear increase in predictions with increasing pH. However, LR, MR, and ENR predictions linearly increase with increasing pH. For season, the partial-dependence profiles show a combination of patterns, increasing or decreasing average predictions across MLs with relatively declining values between LSMA and LAMW, constant values between LAMW and LWMS and increasing values from LWMS to LSES (Fig. 4 (A), Figure S3). The average PDP predictions nonlinearly decreased with conductivity by KNN, XGB, ANN55, GBM, and NNT but nonlinearly increased by ANN64 and ANN10 (Fig. 4 (B), Figure S3). Notably, average predictions linearly or nonlinearly increase in all MLs with increasing temperature until a threshold is reached (Fig. 4 (C), Figure S3). In the same manner, an increase in TSS had a corresponding linear increase in average predictions in LR, MR, KNN and ENR and a nonlinear increase in GBM, NNT, SVR, and M5P until TSS >200 mg/ml, after which a constant/declining prediction began (Fig. 4 (C) and Figure S3). The partial-dependence profiles for salinity showed a declining trend after a constant or an increasing value to 0.4 PSU in RF, SVR, XGB, BRT, and KNN and showed a constant trend in DTR and ENR (Fig. 4 (D), Figure S3). Higher salinity (>0.4 PSU) had an inverse relationship with average predictions. In a pattern similar to temperature, the increase in BOD had a linear or nonlinear increase in the average prediction in all models until an upper limit was reached, where prediction remained constant or started declining (Fig. 4 (D), Figure S3). Additionally, the average predictions of the models were higher at low concentrations of DO but declined with increases in DO after a threshold was reached (Fig. 4 (E) and Figure S3).

## 4. Discussion

The importance of SVs in the incidence, prevalence, and outbreak of infectious diseases and their modelling has been gaining research interest from environmental, clinical, and global change perspectives. The significance of seasonality in the epidemiology of infectious diseases (especially those with high case-fatality rates such as cholera) has been documented in more than 34 countries in Sub-Saharan Africa and nations beyond Africa (Perez-Saez et al. 2022; De Magny et al., 2008; Ruiz-Moreno et al., 2007). In this study, the effects of SVs in environmental fluxes on PDP were modelled.

### 4.1. Site-specific seasonal nature of PEs and PDP

The mean values of PEs such as pH (6.5–8), EC (≤600 μs/cm), TDS (≤500 mg/l), TSS (500 mg/l), BOD (≤6 mg/l), TBS (≤5 NTU), and temperature (≤25–30 °C) across locations in the freshwater study area fell within permissible standard limits, apart from salinity (>0.05 PSU) and TBS (>5 NTU) (WHO, 2008). However, this finding did not imply that the water quality at the sites was suitable for consumption and other purposes. The exceedance of TBS and salinity above the acceptable limit at all sites could be linked to incessant pastoral activities along the freshwaters coupled with resuspension of river sediments. Generally, freshwater quality degradation displays an unprecedentedly higher degree of severity in agricultural land areas and residential, commercial, and pasture environments than in forest areas (Mehaffey et al., 2005). The catchment biogeochemistry, topography, climate, and land use cover also contribute to surface water quality deterioration (Pratt and Chang, 2012). Human activities, run-off, and point source pollution are the major influences that can lead to unacceptable TBS and salinity. Relatively high values of EC, TDS, BOD, and TSS at various points in the freshwater were observed. Most locations were near one or more combinations of informal residential, commercial, pasture, and farmland. The high PDP could be jointly (partly) attributed to high TBS/TSS, which protects the pathogen against direct solar (ultraviolet, UV)

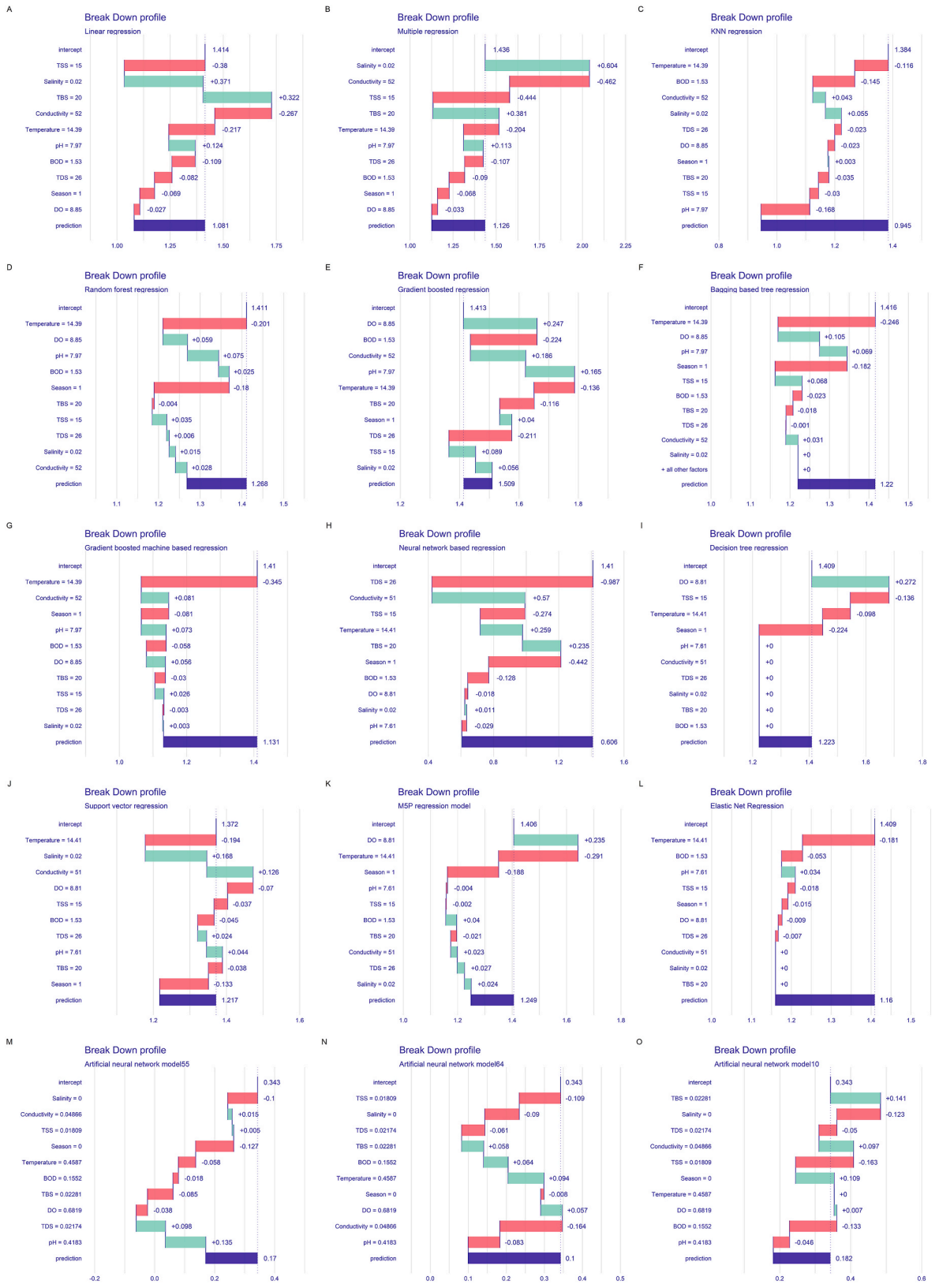


**Fig. 2.** Breakdown plot of seasonal PE contributions in ML models for predicting PDP in the aquatic milieu. Values presented with variables on the y-axis denote the mean value of the corresponding variable required to predict PDP in the aquatic milieu. The values represented by green and red bars denote the minimum flux (dynamic changes/influence) of each variable from its mean value, presented as the ordinate, which will impact the predicted magnitude of PDP from the baseline presented on the plots. Each variable contribution or importance is interpreted in terms of the absolute value of its green or red bar. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

**Table 2**
Comparative predictive performance of eight ML models in predicting the impacts of SVs/PEs on PDP in the aquatic milieu.

| Rank | Model | MSE | RMSE | $R^2$ | MAD |
|---|---|---|---|---|---|
| 1 | XGB | 0.0025 | 0.0501 | 0.9978 | 0.0275 |
| 2 | ANN64 | 0.0059 | 0.0767 | 0.9147 | 0.0467 |
| 3 | ANN10 | 0.0062 | 0.0789 | 0.9098 | 0.0378 |
| 4 | ANN55 | 0.0118 | 0.1087 | 0.8289 | 0.0564 |
| 5 | RF | 0.0316 | 0.1779 | 0.9725 | 0.0935 |
| 6 | M5P | 0.0330 | 0.1817 | 0.9717 | 0.0949 |
| 7 | BRT | 0.1737 | 0.4168 | 0.8491 | 0.2871 |
| 8 | NNT | 0.2448 | 0.4948 | 0.7899 | 0.3038 |
| 9 | KNN | 0.2674 | 0.5171 | 0.7748 | 0.2685 |
| 10 | SVR | 0.2855 | 0.5343 | 0.7550 | 0.2086 |
| 11 | DTR | 0.2924 | 0.5408 | 0.7490 | 0.3030 |
| 12 | GBM | 0.3643 | 0.6035 | 0.6836 | 0.3692 |
| 13 | LR | 0.5207 | 0.7216 | 0.5614 | 0.5260 |
| 14 | ENR | 0.5216 | 0.7222 | 0.5523 | 0.5158 |
| 15 | MR | 0.5245 | 0.7242 | 0.5582 | 0.5468 |

MSE, RMSE, $R^2$ (can be interpreted as the fraction of the total variance of "explained" by the model), and MAD.

irradiation and enhanced nutrients due to point source pollution (BOD) by wastewater, animal wastes, and sediment resuspension by animal activities. High concentrations of PDPs are normally associated with sediment, biofilm surfaces and substrata media.

In addition, nutrient exchanges often occur between water columns and sediments in waterbodies, and thus, nutrient exchanges can exhibit a critical impact in modulating the aquatic ecosystem (Sinkko et al., 2013). Riverbed resuspension increases overlying water nutrient conditions and thereby modulates water PEs. Additionally, the average salinity at Seymour, Fort Beaufort, and StuttGResrv above the acceptable limit indicated possibilities of point source pollution from wastewater treatment plant effluents from nearby wastewater treatment plants (Seymour and Fort Beaufort). The average pH at Balfour, Binfielddam, and StuttFbrid fell within the optimal range of growth requirements for PDP. The high EC at Fort Beaufort and StuttVBridg1 suggests the discharge of wastes or pollutants that were enriched with dissolved ion contents. However, the EC at StuttFbridg may stem from an inflow of dissolved ion content from the agricultural run-off of fertilizers, herbicides, and pesticides from adjoining farmlands, which promotes microbial growth. Agricultural run-offs are usually rich in nitrogen, sulfate, and phosphates. EC has been shown to exhibit a high positive relationship with sulfate, Mn, Zn, and Cd ions in freshwater (Olıas et al., 2004). These ions are important components of many agricultural chemicals and inputs that vary with planting seasons.

Generally, the activities of sediment microorganisms are well documented to drive sediment chemistry, organic deposition, biogeochemical transformation (enhancing EC-component ions) and nutrient exchange involving nitrogen and phosphorus (Hou et al., 2013; Sinkko et al., 2013; Hupfer et al., 2007). Therefore, these microorganisms impact PEs of the overlying water column to different degrees and some nutrient concentrations (Whitman and Nevers, 2003) in addition to riverbed resuspension, which leads to higher TBS and TSS levels. TBS and TSS are known to shield column microorganisms against the lethal activities of solar ultraviolet radiation. The occurrence of *Epistylis* sp.—a prawn pathogen in the aquatic culture system—for example, has been shown to be significantly correlated with TBS (Gutiérrez-Salazar et al., 2011).

High salinity can have a negative impact on PDP, as the organism generally has a low tolerance for high concentrations of salt in freshwater. *Plesiomonas* thrive well under saline conditions $\leq 4‰$ (Janda et al., 2016). However, relatively high PDP with a relatively high level of salinity at some points probably reflects adaptation to a halophilic lifestyle by the bacterium following prolonged or accumulative salinization of the sites due to pollution. Bacteria generally display genome

**Table 3**
Mean variable importance over 100 permutations in predicting the impacts of SVs of PEs on PDP in the aquatic milieu.

| Rank | XGB | RMSE loss | RF | RMSE loss | BRT | RMSE loss | KNN | RMSE lose | GBM | RMSE loss |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Temperature | 0.6491 | Temperature | 0.8166 | Temperature | 0.6179 | Temperature | 0.2803 | Temperature | 0.3343 |
| 2 | Season | 0.5207 | Season | 0.3657 | Season | 0.2525 | Season | 0.1976 | TSS | 0.0888 |
| 3 | TSS | 0.4249 | DO | 0.2576 | TSS | 0.1417 | BOD | 0.1449 | Season | 0.0841 |
| 4 | DO | 0.3018 | TSS | 0.2546 | DO | 0.1030 | pH | 0.1367 | DO | 0.0395 |
| 5 | pH | 0.2890 | pH | 0.2132 | pH | 0.0966 | DO | 0.1153 | BOD | 0.0212 |
| 6 | BOD | 0.2533 | BOD | 0.1632 | BOD | 0.0415 | TBS | 0.0702 | TBS | 0.0107 |
| 7 | TBS | 0.2506 | TBS | 0.1365 | TBS | 0.0409 | TSS | 0.0569 | Conductivity | 0.0054 |
| 8 | Conductivity | 0.2053 | Conductivity | 0.0555 | Conductivity | 0.0188 | Salinity | 0.0224 | pH | 0.0050 |
| 9 | TDS | 0.1572 | TDS | 0.0305 | TDS | 0.0068 | Conductivity | 0.0190 | TDS | 0.0039 |
| 10 | Salinity | 0.0507 | Salinity | 0.0215 | Salinity | 0.0062 | TDS | 0.0087 | Salinity | 0.0005 |
| Rank | NNT | RMSE loss | LR | RMSE loss | MR | RMSE loss | DTR | RMSE loss | SVR | RMSE loss |
| 1 | TDS | 1.0878 | TSS | 1.3168 | TSS | 1.6093 | Temperature | 0.6419 | Temperature | 0.3961 |
| 2 | Conductivity | 1.0474 | TBS | 1.1436 | TBS | 1.4433 | Season | 0.2976 | Season | 0.1703 |
| 3 | TSS | 0.8097 | Temperature | 0.4222 | Salinity | 0.4199 | TSS | 0.1926 | DO | 0.1106 |
| 4 | TBS | 0.7310 | Salinity | 0.1824 | Temperature | 0.3980 | DO | 0.1369 | TDS | 0.0855 |
| 5 | Temperature | 0.5735 | Conductivity | 0.1062 | Conductivity | 0.2833 | pH | 0.0000 | BOD | 0.0836 |
| 6 | Season | 0.3596 | BOD | 0.0390 | DO | 0.0368 | Conductivity | 0.0000 | Conductivity | 0.0812 |
| 7 | DO | 0.1274 | DO | 0.0263 | BOD | 0.0337 | TDS | 0.0000 | Salinity | 0.0800 |
| 8 | BOD | 0.0542 | pH | 0.0169 | TDS | 0.0186 | Salinity | 0.0000 | pH | 0.0755 |
| 9 | pH | 0.0506 | TDS | 0.0113 | pH | 0.0157 | TBS | 0.0000 | TSS | 0.0599 |
| 10 | Salinity | 0.0003 | Season | 0.0034 | Season | 0.0041 | BOD | 0.0000 | TBS | 0.0488 |
| Rank | M5P | RMSE loss | ENR | RMSE loss | ANN55 | RMSE loss | ANN64 | RMSE loss | ANN10 | RMSE loss |
| 1 | Temperature | 0.7990 | TSS | 0.4798 | TSS | 0.2743 | TSS | 0.4821 | TSS | 0.3793 |
| 2 | Season | 0.4328 | Temperature | 0.4113 | Salinity | 0.2447 | TBS | 0.4758 | TBS | 0.3311 |
| 3 | DO | 0.2814 | TBS | 0.3489 | TBS | 0.2293 | Salinity | 0.4264 | Season | 0.2370 |
| 4 | TSS | 0.2150 | Salinity | 0.1813 | Conductivity | 0.2116 | Conductivity | 0.3840 | Temperature | 0.1905 |
| 5 | pH | 0.2023 | Conductivity | 0.1222 | Season | 0.2042 | TDS | 0.2573 | Conductivity | 0.1768 |
| 6 | BOD | 0.1343 | pH | 0.0277 | Temperature | 0.1389 | Season | 0.2035 | Salinity | 0.1677 |
| 7 | TBS | 0.1006 | DO | 0.0242 | pH | 0.0926 | Temperature | 0.1916 | TDS | 0.1626 |
| 8 | Conductivity | 0.0568 | BOD | 0.0168 | DO | 0.0842 | pH | 0.1506 | pH | 0.1447 |
| 9 | TDS | 0.0381 | TDS | 0.0165 | BOD | 0.0746 | BOD | 0.1221 | BOD | 0.0711 |
| 10 | Salinity | 0.0205 | Season | 0.0041 | TDS | 0.0511 | DO | 0.0700 | DO | 0.0526 |

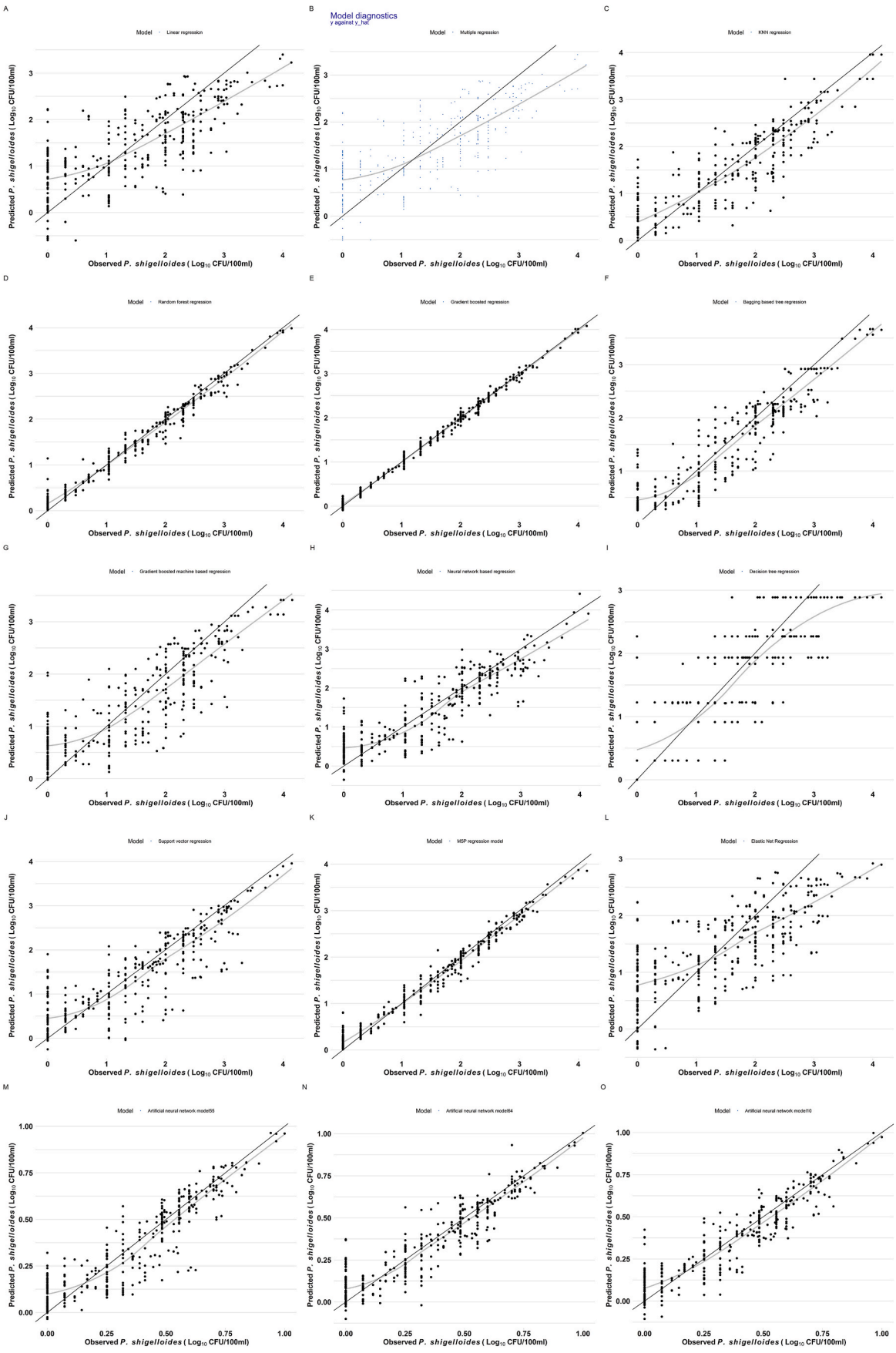


**Fig. 3.** Comparison of observed and predicted PDP values by the fifteen models.

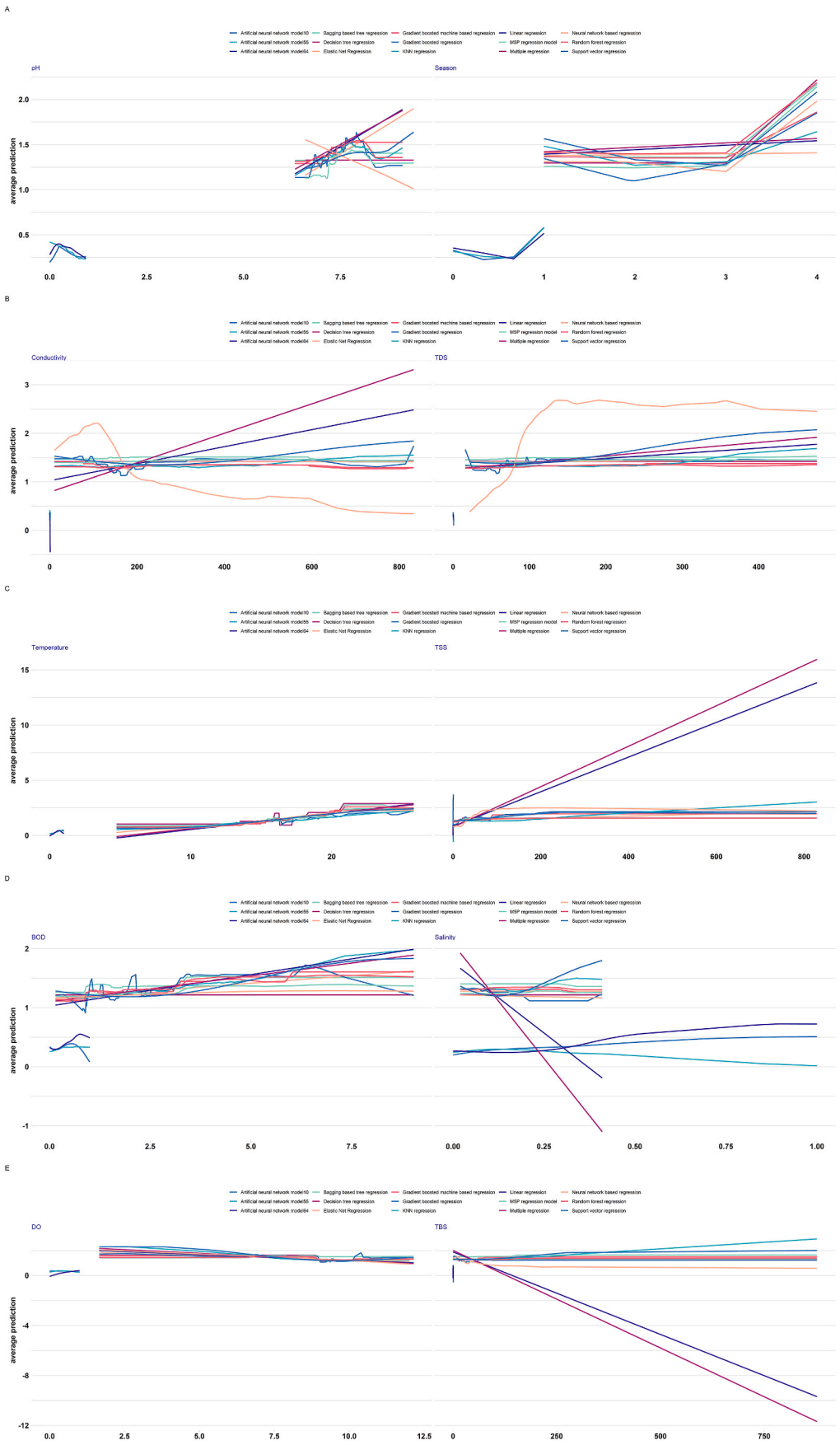


**Fig. 4.** Contrastive partial-dependence profiles for the fifteen models and predictive variables on PDP.

plasticity, which enables them to adapt to harsh conditions.

The DO level was higher at some locations because of the high flow rate, turbulence, mixing conditions, and reaeration. However, a high level of DO is problematic for facultative anaerobic bacteria and PDP.

### 4.2. Seasonal nature of PEs and PDP in the rivers

The PEs and PDP along the river courses varied across seasons and were suggestive of seasonal effects of neutral and niche-related processes in the freshwater. Niche-associated processes (Cavender-Bares et al., 2009) involved selection caused by microbial interactions (facilitation, competition, predation, and mutualism) and environmental filtering (abiotic environment), whereas neutral effects (Hubbell, 2001) consist of probabilistic dispersal, unpredictable disturbances, and stochastic processes, including birth/death events (Hanson et al., 2012; Zhou et al., 2013; Ren et al., 2015), due to seasonality. Seasonality affects microbial PD, as different seasons are accompanied by different activation and inactivation of various molecular switches (circadian rhythms) to enable microbial survival of the imminent change or consequences. Additionally, SVs in the hydrologic cycle of the freshwater system can modulate PEs and eventually impact pathogen concentrations. The positive correlation between PDP and pH has been reported elsewhere (Bülger, 2004).

The observed EC and PDP at the Tyhume River in seasons A, B, and C and in the Kat River in D are probably attributed to the effect of SV on the biogeochemistry and mobility of EC-determining components at the sites. Olıas and coworkers (2004) recorded the highest EC values during September and October in the Odiel River, Spain. Additionally, the seasonal TDS, salinity, and PDP were solely predominant in the Tyhume River for seasons A, B, and C. This finding could be attributed to ionic pollution from farm activities and other sources. However, the PDP and salinity in LSES (season D) in the Kat River suggest that the period was characterized by high salinity, which was detrimental to *P. shigelloides* growth. LSES was characterized by a high degree of land preparation and other farm management practices for the summer rainy season in the Eastern Cape Province (ECP), South Africa. Uncertain pH values and low salinity gradients in freshwater are unique to the rainy season and may trigger the induction or expression of microbial virulence genes, and in part, explain the seasonality of infection or disease outbreaks (Prayitno and Latchford, 1995). Higher (intermediate) salinity gradients have been shown to increase some microbial community doubling times in coastal rivers from 1 to 3 days in summer and to $>$ 4 days in spring (Crump et al., 2004).

The effect of seasonality on the degree of dependency of PDP on temperature was significant in all the rivers, except for the Kubusie River for seasons A, B, and C. Higher temperatures are usually characteristics of the wet season in South Africa (Abia et al., 2015). These observations parallel the seasonality of *Plesiomonas* infection and differences in disease prevalence that are generally associated with temperate and tropical/subtropical climates (Stock, 2004; Jagger, 2000; Miller and Koburger, 1985). Elevated freshwater temperatures may produce a higher growth rate and water column concentrations of *Plesiomonas* (Stock, 2004). The present study's results show agreement with those of Bülger (2004), who observed a significant positive correlation between *Plesiomonas* density and temperature in Nilufer Stream, Bursa, Turkey.

In this study, significant levels of TSS/TBS were observed in the Kat River from February to July (seasons A and B) and TBS and TSS in the Tyhume River and Kat River, respectively, in August–October (season C). The seasonal differences in TBS/TSS in the rivers could be linked to the seasonal differences between human-induced processes and natural processes. For instance, seasonal activities/phenomena affect riverbed resuspension. During the dry winter season, farmers are observed walking their animals (cattle, sheep and goats) to the rivers for direct watering. Livestock crossings (cattle crossings in particular) of rivers have been reported to result in the resuspension of riverbeds, leading to higher TBS and sediment-borne microorganisms in the water column (Abia et al., 2017). The role of cows crossing on the *Escherichia coli* concentration in overlying water has been reported elsewhere (McDaniel et al., 2013). Sediment raking was shown to increase the *E. coli* concentration 7.9 times higher than that in undisturbed sediment in the Apies River and positively correlated TBS and the *E. coli* concentration (Abia et al., 2017). TBS and TSS generally increase in freshwater during the rainy season due to run-off from immediate farms and urban areas (Martinez et al., 2014; Ramos et al., 2006) and to riverbed resuspension as a result of increased flow velocity (Abia et al., 2017). Additionally, microbial PD is known to be correlated with TBS in water, as most bacteria are attached to suspended solids (Taylor et al., 1993). High TBS has been associated with increased TSS from stormwater events after heavy rainfall, leading to a higher density of *E. coli* in the overlying water column in the streams (Craig et al., 2004). In addition, storm events increase the PD of indicator microorganisms by 30–55% compared with 20–35% during normal river flow and are also linked to suspended particles within the aquatic milieu (Krometis et al., 2007; Characklis et al., 2005).

PDP and BOD displayed a significant trend throughout the seasons in the three freshwaters, apart from seasons A, B and C (February–October) for the Kubusie River. The availability of nutrients (BOD) is strongly related to PDP (Rippey and Cabelli, 1980; Bülger, 2004). In general, this study parallels the seasonal prevalence and distribution of *Plesiomonas* infections. Mahmoud (2016) obtained more positive isolation of *Plesiomonas* during spring and summer periods compared with other seasons in aquarium water and diarrhoeal samples.

### 4.3. ML predicted PDP values and model-specific contributory variables

DTR had the highest model-predicted PDP among the ML algorithms, followed by XGB and the GBM (Tables 1–3). However, XGB outperformed all the models in all regression metrics. The results show agreement with previous studies that found XGB to outperform other models in predicting and tracking microbial sources (Wu et al., 2021). Although DTR and XGB identified temperature, season, TSS, and DO in decreasing order as important factors in predicting PDP, while XGB still captured the relationship between PDP and other variables (pH, BOD, TBS, conductivity, TDS, and salinity) and incorporated it in its prediction, DTR did not yield the same result (Table 3). This discrepancy might be responsible for the poor performance of DTR in terms of the regression metrics. Similarly, gradient boosting regression (GBR) has been shown to have the highest $R^2$ value and lowest MSE value compared with DTR, SVR, ANN, kernel ridge regression, and Bayesian ridge regression in forecasting the crack closure percentage (CCP) of bacteria-based, self-healing bioconcrete (Zhuang and Zhou, 2019). DTR has also been shown to outperform MR in predicting potential airborne bacterial hazards and $PM_{10}$ concentrations during Asian dust events (Yoo et al., 2018). However, both DTR and GBR have been reported to achieve better predictive performance in determining the CCP of the bacteria-based, self-healing bioconcrete (Zhuang and Zhou, 2019), as observed in this study. DTR is not as common in the literature in addressing microbial-related regression problems. However, the poor regression metrics of DTR could possibly be improved with more data than the limited dataset fitted in this study.

The PDP prediction by models followed decreased progression from GBM $>$ BRT $>$ RF $>$ M5P $>$ SVR. The various performances of the different ML algorithms can be linked with the differential model's ability to effectively identify the complex relationship between the input variables and PDP. Gradient boosting classifiers are known to outperform ridge and multivariable logistic regressions in predicting microorganisms of lower taxonomic levels in microbiome datasets (Liu et al., 2022). While the RF predicts PDP better than M5P and SVR, RF-based regression is susceptible to outliers' influence (i.e., extremely large or small values) when present in a dataset and may lead to underestimation (large value) or overestimation (small value) (Zhang and Lu, 2012). LR and MR had higher model-predicted PDPs values than ENR, KNN, NNT,

ANN10, ANN64, and ANN55. However, LR and MR yielded poor regression performance metrics, which might indicate that the models were inadequate and ineffective in capturing the complex relationship between seasonal interactions of PEs and PDP, unlike other models (Tables 1–3). The predictive performance of the models depends on their ability to identify patterns/relationships that were not yet known or understood between PDP and seasonal PEs.

The findings of this study revealed varying degrees of seasonal PE contributions to the predictive power of ML algorithms (Fig. 3). While TSS, salinity, and TBS contributed in descending order to LR model-predicted PDPs, in a similar pattern, salinity, EC, TSS, and TBS contributed to MR model-predicted PDPs. This finding implied that both LR and MR captured linear interactions between PDP and nutrient loading of the rivers. However, LR and MR are inherently incapable of modelling nonlinear relationships. TSS and TBS are major components derived from nutrient loads of freshwater and nutrient availability is strongly related to PDP (Rippey and Cabelli, 1980; Bülger, 2004), a relationship which might to be a linear association. KNN identified pH, BOD, and temperature as the best 3 variables, whereas RF identified temperature, DO, and pH as the best 3 variables. The observed differences in the factors with the highest contributions to PDP prediction by XGB (DO, BOD), BRT (temperature, TSS), GBM (temperature, DO), NNT (TDS, conductivity), DTR (DO, TSS), SVR (temperature, salinity), M5P (temperature, DO), ENR (salinity, conductivity), ANN64 (TDS, temperature, ANN55 (pH, season), and ANN10 (season, TSS) could be attributed to the models' differential abilities to capture more than one network of complex and/or nonlinear relationships between the variables and PDP. It appeared that these models captured more than one network of latent variables influencing PDP in aquatic environments, previous studies as confirmed the ability of the ML models to improve approximation, continuity, smoothness, extrapolation, and interpolation of nonlinear relationships (Kappler et al., 2005; Ao et al., 2019).

### 4.4. Regression metric-based performance of the MLI predictive model for PDP

The assessment of the performance and accuracy of the MLI predictive model for seasonal PDP in aquatic milieu identified XGB, ANN64, ANN10, ANN5, RF, and M5P in the 1st, 2nd, 3rd, 4th, 5th, and 6th positions, respectively (Table 2). NNT, KNN, SVR, DTR, GBM, LR, ENR, and MR had MSE $\geq$0.2448, RMSE $\geq$0.4948, R2 $\leq$0.7899, and MAD$\geq$0.3038 values, which revealed that they underperformed in predicting PDP relative to other ML algorithms (Table 2). Our results are consistent with previous studies. XGB predicted *Helicobacter pylori* infection status more accurately than KNN, logistic regression with Lasso penalty, SVM, RF, and naive Bayes in their classification-based models (Tran et al., 2022). XGB similarly outperformed Elastic Net, RF, KNN, SVM, and light gradient boosting in terms of accuracy in predicting HIV status based on sociobehavioural-driven data (Mutai et al., 2021). The RF has been previously ranked second to a DNN (ANN) in terms of the predictive accuracy of *Vibrio* pathogen abundance on microplastics in mariculture zones and estuaries, whereas **XGB** had poor performance (Jiang et al., 2022). In a study by Jiang et al. (2022), the ANN outperformed SVR, Elastic Net, RF, and XGBoost in predicting microplastic-associated *Vibrio* pathogens. Furthermore, Wu et al. (2021) discovered that XGB, RF, and KNN ranked first, second, and third, respectively, in performance and the average accuracy in predicting and tracking microbial sources. Notwithstanding that all ANN-based models had good regression metrics and performance (Table 2), they had poor predicted values of PDP (Table 1). This finding may be attributed to the limited dataset applied in training the models in the present study. The ANN is a deep learning algorithm whose performance largely depends on the volume of the dataset explored in training it. A well-trained ANN model has been shown to rapidly estimate real-time, airborne fungal concentrations and to provide ultrafast estimation with acceptable accuracies (Liu et al., 2018). Additionally, GBM, SVR, and the RF had been previously demonstrated as robust models in predicting microbial biohydrogen production from wastewater with $R^2$ values $\geq$ 0.893 and MSE values $\leq$ 0.016 (Hosseinzadeh et al., 2022); only RF had robust $R^2$ values (0.97) and MSE values (0.0316) for predicting PDP in this study. In the prediction of *Fusarium culmorum* and *F. proliferatum* growth rates and their ability to produce mycotoxins, the RF was reported to outperform NNT models and XGB (Srinivasan et al., 2022). SVR has also been reported to outperform Gaussian process regression, extremely randomized tree regression, and the traditional Baranyi model in terms of regression metrics in predicting *Escherichia coli* O157 growth behaviour (Koyama et al., 2022).

Generally, the better predictive performance of a model is connected to low MSE, RMSE, and MAD values but not a higher $R^2$ value (Biecek and Burzykowski, 2021). For a "perfect" model, which exactly predicts (fits) all, MSE = 0, RMSE = 0, MAD = 0, and $R^2$ = 1 (Biecek and Burzykowski, 2021). XGB has merits over other models because of its ensemble nature, similar to RF, a multiple decision tree-based inference system, and its ability to avoid overfitting via additional regularization, thereby reducing prediction errors (Wu et al., 2022; Wang et al., 2020). In addition, the performance of RF models is largely dependent on dataset volume and the ensemble of their trees (Gupta et al., 2021).

### 4.5. Importance variables for the MLI predictive model for PDPs

The feature importance selection of the variables in this study identified temperature as the major relevant factor in predicting PDP in XGB, RF, BRT, KNN, GBM, DTR, SVR, and M5P. These results were corroborated by a previous study on *V. parahaemolyticus* that identified sea surface temperature as the first and key input parameter for forecasting microorganism abundance in seawater (DeLuca et al., 2020). Temperature generally has a significant influence on the physiological growth of *P. shigelloides* (Janda et al., 2016). However, temperature alone cannot effectively predict PDP due to complex interactions between PDP and environmental and anthropogenic factors. For this reason, the RF in a recent study indicated that water temperature could not exclusively justify the differences in *Salmonella* detection among seasons, as it was governed by a network of complex interactions that might influence the likelihood of seasonal and regional detection of *Salmonella* (Toro et al., 2022). Additionally, TSS ranked first in important variables for forecasting PDP by LR, MR, ENR, ANN55, ANN64, and ANN10. Microplastics are a notable example of emerging TSS pollution that has been shown to greatly increase the relative abundance of pathogens and antimicrobial resistance genes in aquatic environments compared with corresponding sediments and seawater (Jiang et al., 2022; Wang et al., 2021). In addition, TSS influences the water temperature, serving as an insulator/storage to prevent heat loss in the environment. This influence further highlighted the possibility that ML models may even be able to capture relationships between microorganisms and environmental variables that are not yet understood or known. An ensemble prediction system from simple, complex and/or combinations of ML modelling can account for faulty detection of microbial pathogens in the aquatic milieu (MacKenzie et al., 2002) and increase skill predictions of hidden (nonlinear) interactions between microbes and variables in datasets (DeLuca et al., 2020).

Season was identified as the second most important factor in predicting PDPs by 7 (XGB, RF, BRT, KNN, DTR, SVR, and M5P) of the 15 ML models, whereas TBS ranked second based on the results of LR, MR, ANN64, and ANN10. The role of season in the prediction of PDP could be attributed to the SV in temperature and other PEs coupled with anthropogenic activities that may increase the pollution/nutrient loadings of freshwater. Similarly, Toro et al. (2022) reported that the RF identified season/month as one of the most influential factors in forecasting *Salmonella* detection in aquatic environments. However, the true picture of seasonal factors in predicting microorganisms was only evident when samples are collected across seasons. Our data satisfied this condition compared with Zhou et al. (2022), whose data were

collected during the wet season only, and thus, their RF models predicted metal resistance gene abundance in estuaries in the wet season only (Zhou et al., 2022). The overall results suggest that temperature, season, and TSS are essential factors in most ML model predictions of PDPs in freshwater.

### 4.6. Sensitivity of the models to input variables

The results from residual-diagnostic plots showed that XGB and the RF fit the PDP closest to the observed values with approximately overlapping smoothed trends in XGB (Fig. 3). This finding implied that XGB estimates the actual value of PDP better than other models. Additionally, the observed smoothed trends of most models above or below the straight line at higher values or lower values, respectively, imply that the models overestimate or underestimate the actual PDP at low density and higher density, respectively.

The summary of the results in Fig. 4 (Figure S3) implied that PDP had an increasing relationship with unit increases in pH until the threshold $\geq$ pH 8.0 was reached. This result is consistent with a previous report that *P. shigelloides* has an optimal pH range of 4.5–9 (Janda et al., 2016).

The findings in this study identified SVs that affected the average PDP predictions and decreased values between LSMA and LAMW and increased values from LWMS to LSES. This result was further corroborated by the increase in average predictions with increasing temperature until the upper bound of the temperature requirement. Additionally, the average prediction of PDPs was nonmonotonously sensitive to increases in TSS in GBM, NNT, SVR, and M5P and linearly sensitive in LR, MR, KNN and ENR until the upper limit TSS >200 mg/ml was reached. Although there is no specific standard TSS requirement for *P. shigelloides* growth, these results suggest that excessive TSS pollution of a toxic nature might be problematic for *P. shigelloides* growth. The results for salinity showed a negative trend between PDPs and salinity at higher salinities (>0.4 PSU). *P. shigelloides* is known to thrive in salinities between 0% and 4% (Janda et al., 2016). While an increase in BOD led to an increase in the average prediction of PDP until an upper threshold beyond the growth requirement, low concentrations of DO favoured higher PDP.

The present study had some limitations. The first limitation is the inability to fit more deep learning models that could potentially outperform the best model identified in this study in predicting PDP due to the available data. Second, the lack of spatial inputs in the model is another shortcoming, as the potential spatial distributions of the pathogen in the water systems vary. However, this study investigated for the first time the SV in PDP in the aquatic milieu and demonstrated that ML algorithms are promising tools to accurately predict PDP in rivers, provide a short turnaround time, reduce labour of laboratory experiments, reduce monitoring parameters, and provide a framework for decisive, proactive decision-making. In addition, ML models outperformed traditional mathematical models (Long et al., 2022), as they are capable of modelling complex, nonlinear and multidimensional interactions and relationships between PDPs and PEs as well as their modulating/underlying anthropogenic events. ML models can robustly fit multiple data types and capture relationships between them, unlike traditional mathematical models. MLI predictive models are easily tuneable (e.g., with regional data), adaptable, implementable, and deployable on any platform irrespective of the origin of development.

It is recommended that extensive data and some deep learning ML algorithms be explored in future studies to build a transfer learning framework for PDP. A web-based or mobile interphase application could be further deployed to predict PDP in the aquatic milieu with few significant variables (temperature, season, and TSS) identified.

### 4.7. Conclusions

The present investigation showed that SVs significantly pattern PDP in the aquatic milieu, exhibiting site-specific associations with PEs. MLI had varied performance in forecasting PDP in aquatic environments, with **XGB** (MSE = 0.0025, RMSE = 0.0501; $R^{2=}$0.998, MAD = 0.0275) **outperforming** other predictive models in all regression metrics**.** The overall results of the MLI models revealed that temperature, TSS, and season were primary factors influencing and essential to accurately predict PDP by most of the MLIs. The ML predictive modelling technique could promisingly be exploited to complement laboratory-based monitoring of PDP and the PD of other pathogens, especially in low-resource settings in response to seasonal fluxes, and to provide insights into the potential public health risks of emerging pathogens and TSS pollution (e.g., nanoparticles, micro, and nanoplastics) in aquatic systems. The model outputs can provide low-cost, short turnaround time, and effective early warning information to assist watershed managers in making appropriate decisions about water resource protection and sustainable public health protections. The current investigation signified that MLI-based techniques can achieve extremely fast estimation of the abundance of PDP with significant accuracy. Additionally, fish farmers could benefit from the implementation of MLI-based decision systems for PDP in aquaculture management to prevent and avoid gruesome fish mortality and economic loss due to *P. shigelloides* infections and related pathogens. Future studies should involve the input of data from different sources, including aquaculture systems and relevant stakeholders, in considering suitable, highly sensitive/or low-specificity and meritorious MLI models for predicting PDP in aquatic systems for early warning risk assessments and preventions to account for potential heterogeneities in different settings and systems.

## Author contributions

Conceptualization: T.C.E., A.I.O.; Investigation, Software and Formal analysis: T.C.E.; Resources: A.I.O; Writing - original draft preparation and interpretations: T.C.E.; A.I. O.; E.O. I.; O.A.I.; Supervision: A.I.O.; Funding acquisition: A.I.O.; critical review for intellectual contents: T.C.E.; A.I. O.; E.O. I.; O.A.I.; All authors contributed to writing - review and editing, and approved the final version of the manuscript for publication.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgements

We are grateful to the South African Medical Research Council (UFH/P790), the National Research Foundation, The World Academy of Science (Grant Numbers: 99796 and 116382), the University of Sharjah, the African-German Network of Excellence in Science, the Federal Ministry of Education and Research and the Alexander von Humboldt Foundation for their financial support.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envpol.2022.120734.

## References

Abbott, S.L., Kokka, R.P., Janda, J.M., 1991. Laboratory investigations on the low pathogenic potential of *Plesiomonas shigelloides*. J. Clin. Microbiol. 29 (1), 148–153.
Abia, A.L.K., James, C., Ubomba-Jaswa, E., Benteke Momba, M.N., 2017. Microbial remobilisation on riverbed sediment disturbance in experimental flumes and a

human-impacted river: implication for water resource management and public health in developing sub-Saharan African countries. Int. J. Environ. Res. Publ. Health 14 (3), 306.
Abia, A.L.K., Ubomba-Jaswa, E., Momba, M.N.B., 2015. Impact of seasonal variation on *Escherichia coli* concentrations in the riverbed sediments in the Apies River, South Africa. Sci. Total Environ. 537, 462–469.
Amini, M.H., Arab, M., Faramarz, M.G., Ghazikhani, A., Gheibi, M., 2021. Presenting a soft sensor for monitoring and controlling well health and pump performance using machine learning, statistical analysis, and Petri net modeling. Environ. Sci. Pollut. Control Ser. 1–17.
Ao, Y., Li, H., Zhu, L., Ali, S., Yang, Z., 2019. The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. J. Petrol. Sci. Eng. 174, 776–789.
APHA (American Public Health Association), 2005. Standard Methods for the Examination of Water and Wastewater, twenty-first ed. American Public Health Association, American Water Works Association, Water Environment Federation, Washington DC http://www.worldcat.org/oclc/156744115/editions?edition sView=true&referer=di.
Band, S.S., Janizadeh, S., Pal, S.C., Chowdhuri, I., Siabi, Z., Norouzi, A., Melesse, A.M., Shokri, M., Mosavi, A., 2020. Comparative analysis of artificial intelligence models for accurate estimation of groundwater nitrate concentration. Sensors 20 (20), 5763.
Behera, B.K., Bera, A.K., Paria, P., Das, A., Parida, P.K., Kumari, S., Bhowmick, S., Das, B. K., 2018. Identification and pathogenicity of *Plesiomonas shigelloides* in silver carp. Aquaculture 493, 314–318.
Biecek, P., Burzykowski, T., 2021. Explanatory Model Analysis: Explore, Explain and Examine Predictive Models. Chapman and Hall/CRC.
Bonadonna, L., Briancesco, R., Ottaviani, M., Veschetti, E., 2002. Occurrence of *Cryptosporidium* oocysts in sewage effluents and correlation with microbial, chemical and physical water variables. Environ. Monit. Assess. 75 (3), 241–252.
Bülger, B., 2004. Occurrence of *Plesiomonas shigelloides* and relationship with faecal pollution in nilufer stream, Bursa, Turkey. Turkish Electron J Biotechnol 2, 22–29.
Cavender-Bares, J., Kozak, K.H., Fine, P.V., Kembel, S.W., 2009. The merging of community ecology and phylogenetic biology. Ecol. Lett. 12 (7), 693–715.
Centers for Disease Control and Prevention, 2015. Diarrhoea: common illness, global killer. Available online: http://www.cdc.gov/healthywater/global/diarrhea-burden.html. (Accessed 20 April 2019).
Characklis, G.W., Dilts, M.J., Simmons III, O.D., Likirdopulos, C.A., Krometis, L.A.H., Sobsey, M.D., 2005. Microbial partitioning to settleable particles in stormwater. Water Res. 39 (9), 1773–1782.
Chen, H., Zhao, Y., Chen, K., Wei, Y., Luo, H., Li, Y., Liu, F., Zhu, Z., Hu, W., Luo, D., 2022. Isolation, identification, and investigation of pathogenic bacteria from Common Carp (*Cyprinus carpio*) naturally infected with *Plesiomonas shigelloides*. Front. Immunol. 3292.
Chen, W., Li, Y., Tsangaratos, P., Shahabi, H., Ilia, I., Xue, W., Bian, H., 2020. Groundwater spring potential mapping using artificial intelligence approach based on kernel logistic regression, random forest, and alternating decision tree models. Appl. Sci. 10 (2), 425.
Chen, X., Chen, Y., Yang, Q., Kong, H., Yu, F., Han, D., Zheng, S., Cui, D., Li, L., 2013. *Plesiomonas shigelloides* infection in Southeast China. PLoS One 8 (11), e77877.
Ciznar, I., González-Rey, C., Krovacek, K., Hostacka, A., 2006. Plesiomonas shigelloides. Food-Borne Pathogens: Methods and Protocols 73.
Craig, D.L., Fallowfield, H.J., Cromar, N.J., 2004. Use of microcosms to determine persistence of *Escherichia coli* in recreational coastal water and sediment and validation with *in situ* measurements. J. Appl. Microbiol. 96 (5), 922–930.
Crump, B.C., Hopkinson, C.S., Sogin, M.L., Hobbie, J.E., 2004. Microbial biogeography along an estuarine salinity gradient: combined influences of bacterial growth and residence time. Appl. Environ. Microbiol. 70 (3), 1494–1505.
De Magny, G.C., Murtugudde, R., Sapiano, M.R., Nizam, A., Brown, C.W., Busalacchi, A. J., Yunus, M., Nair, G.B., Gil, A.I., Lanata, C.F., Calkins, J., 2008. Environmental signatures associated with cholera epidemics. Proc. Natl. Acad. Sci. USA 105 (46), 17676–17681.
DeLuca, N.M., Zaitchik, B.F., Guikema, S.D., Jacobs, J.M., Davis, B.J., Curriero, F.C., 2020. Evaluation of remotely sensed prediction and forecast models for *Vibrio parahaemolyticus* in the Chesapeake Bay. Rem. Sens. Environ. 250, 112016.
Ekundayo, T.C., Okoh, A.I., 2019. Modelling the effects of physicochemical variables and anthropogenic activities as ecological drivers of *Plesiomonas shigelloides* distribution and freshwaters quality. Sci. Total Environ. 682, 765–778.
Gardner, S.E., Fowlston, S.E., George, W.L., 1987. In vitro production of cholera toxin-like activity by *Plesiomonas shigelloides*. JID (J. Infect. Dis.) 156 (5), 720–722.
Gupta, S., Aga, D., Pruden, A., Zhang, L., Vikesland, P., 2021. Data analytics for environmental science and engineering research. Environ. Sci. Technol. 55 (16), 10895–10907.
Gutiérrez-Salazar, G.J., Molina-Garza, Z.J., Hernández-Acosta, M., García-Salas, J.A., Mercado-Hernández, R., Galaviz-Silva, L., 2011. Pathogens in Pacific white shrimp (*Litopenaeus vannamei* Boone, 1931) and their relationship with physicochemical parameters in three different culture systems in Tamaulipas, Mexico. Aquaculture 321 (1–2), 34–40.
Hansen, L.K., 1993. Stochastic linear learning: exact test and training error averages. Neural Network. 6 (3), 393–396.
Hanson, C.A., Fuhrman, J.A., Horner-Devine, M.C., Martiny, J.B., 2012. Beyond biogeographic patterns: processes shaping the microbial landscape. Nat. Rev. Microbiol. 10 (7), 497.
Hosseinzadeh, A., Zhou, J.L., Altaee, A., Li, D., 2022. Machine learning modeling and analysis of biohydrogen production from wastewater by dark fermentation process. Bioresour. Technol. 343, 126111.
Hou, J., Song, C., Cao, X., Zhou, Y., 2013. Shifts between ammonia-oxidizing bacteria and archaea in relation to nitrification potential across trophic gradients in two large Chinese lakes (Lake Taihu and Lake Chaohu). Water Res. 47 (7), 2285–2296.
Hu, Q., Lin, Q., Shi, C., Fu, X., Li, N., Liu, L., Wu, S., 2014. Isolation and identification of a pathogenic *Plesiomonas shigelloides* from diseased grass carp. Wei sheng wu xue bao= Acta microbiologica Sinica 54 (2), 229–235.
Hubbell, S.P., 2001. The Unified Neutral Theory of Biodiversity and Biogeography. Princeton University Press, Princeton, NJ.
Hupfer, M., Gloess, S., Grossart, H.P., 2007. Polyphosphate-accumulating microorganisms in aquatic sediments. Aquat. Microb. Ecol. 47 (3), 299–311.
Jagger, T.D., 2000. *Plesiomonas shigelloides* - a veterinary perspective. Infect. Dis. Rev. 2, 199–210.
Janda, J.M., Abbott, S.L., 1993. Expression of hemolytic activity by *Plesiomonas shigelloides*. J. Clin. Microbiol. 31 (5), 1206–1208.
Janda, J.M., Abbott, S.L., McIver, C.J., 2016. *Plesiomonas shigelloides* revisited. Clin. Microbiol. Rev. 29 (2), 349–374.
Jiang, J., Zhou, H., Zhang, T., Yao, C., Du, D., Zhao, L., Cai, W., Che, L., Cao, Z., Wu, X.E., 2022. Machine learning to predict dynamic changes of pathogenic Vibrio spp. abundance on microplastics in marine environment. Environ. Pollut. 305, 119257.
Jiang, J.Z., Liu, Y., Yan, L.H., Yan, Q.G., Wen, X.T., Cao, S.J., Huang, Y., Huang, X.B., Ma, X.P., Han, X.F., Zhao, Q., 2021. Identification and pathogenicity of *Plesiomonas shigelloides* from *Acipenser dabryanus* in China. Aquacult. Res. 52 (5), 2286–2293.
Kampstra, P., 2008. Beanplot: a boxplot alternative for visual comparison of distributions. Journal of Statistical Software, Code Snippets 28 (1), 1–9. http ://www.jstatsoft.org/v28/c01/.
Kappler, K., Kuzma, H.A., Rector, J.W., 2005. A comparison of standard inversion, neural networks and support vector machines. In: Seg Technical Program Expanded Abstracts 2005. Society of Exploration Geophysicists, pp. 1725–1727.
Kassambara, A., 2018. Ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.1.7. https://CRAN.R-project.org/package=ggpubr.
Keating, J.P., 2005. Chronic diarrhea. Pediatr. Rev. 26 (1), 5.
Kombo, O.H., Kumaran, S., Sheikh, Y.H., Bovim, A., Jayavel, K., 2020. Long-term groundwater level prediction model based on hybrid KNN-RF technique. Hydrology 7 (3), 59.
Koyama, K., Kubo, K., Hiura, S., Koseki, S., 2022. Is skipping the definition of primary and secondary models possible? Prediction of Escherichia coli O157 growth by machine learning. J. Microbiol. Methods 192, 106366.
Krometis, L.A.H., Characklis, G.W., Simmons III, O.D., Dilts, M.J., Likirdopulos, C.A., Sobsey, M.D., 2007. Intra-storm variability in microbial partitioning and microbial loading rates. Water Res. 41 (2), 506–516.
Lee, J.H., Huang, Y., Dickman, M., Jayawardena, A.W., 2003. Neural network modelling of coastal algal blooms. Ecol. Model. 159 (2–3), 179–201.
Lin, Y.C., Salleb-Aouissi, A., Hooven, T.A., 2022. Interpretable prediction of necrotizing enterocolitis from machine learning analysis of premature infant stool microbiota. BMC Bioinf. 23 (1), 1–29.
Liu, Y., Méric, G., Havulinna, A.S., Teo, S.M., Åberg, F., Ruuskanen, M., Sanders, J., Zhu, Q., Tripathi, A., Verspoor, K., Cheng, S., 2022. Early prediction of incident liver disease using conventional risk factors and gut-microbiome-augmented gradient boosting. Cell Metabol. 34 (5), 719–730.
Liu, Z., Cheng, K., Li, H., Cao, G., Wu, D., Shi, Y., 2018. Exploring the potential relationship between indoor air quality and the concentration of airborne culturable fungi: a combined experimental and neural network modeling study. Environ. Sci. Pollut. Control Ser. 25 (4), 3510–3517.
Liu, Z., Ke, X., Lu, M., Gao, F., Cao, J., Zhu, H., Wang, M., 2015. Identification and pathological observation of a pathogenic *Plesiomonas shigelloides* strain isolated from cultured tilapia (*Oreochromis niloticus*). Wei sheng wu xue bao= Acta microbiologica Sinica 55 (1), 96–106.
Long, B., Fischer, B., Zeng, Y., Amerigian, Z., Li, Q., Bryant, H., Li, M., Dai, S.Y., Yuan, J. S., 2022. Machine learning-informed and synthetic biology-enabled semi-continuous algal cultivation to unleash renewable fuel productivity. Nat. Commun. 13 (1), 1–11.
Mahmoud, W.S., 2016. The prevalence of *Plesiomonas shigelloides* among hospitalized and out-clinic diarrheal patients and the role of the aquarium as a source of infection. Tikrit Medical Journal 21 (1).
Martinez, G., Pachepsky, Y.A., Whelan, G., Yakirevich, A.M., Guber, A., Gish, T.J., 2014. Rainfall-induced fecal indicator organisms transport from manured fields: model sensitivity analysis. Environ. Int. 63, 121–129.
McDaniel, R.L., Soupir, M.L., Tuttle, R.B., Cervantes, A.E., 2013. Release, dispersion, and resuspension of *Escherichia coli* from direct fecal deposits under controlled flows. JAWRA Journal of the American Water Resources Association 49 (2), 319–327.
Mehaffey, M.H., Nash, M.S., Wade, T.G., Ebert, D.W., Jones, K.B., Rager, A., 2005. Linking land cover and water quality in New York City's water supply watersheds. Environ. Monit. Assess. 107 (1–3), 29–44.
Miller, M.L., Koburger, J.A., 1985. *Plesiomonas shigelloides*: an opportunistic food and waterborne pathogen. J. Food Protect. 48 (5), 449–457.
Mutai, C.K., McSharry, P.E., Ngaruye, I., Musabanganji, E., 2021. Use of machine learning techniques to identify HIV predictors for screening in sub-Saharan Africa. BMC Med. Res. Methodol. 21 (1), 1–11.
Namkung, J., 2020. Machine learning methods for microbiome studies. J. Microbiol. 58 (3), 206–216.
Obi, C.L., Potgieter, N., Bessong, P.O., Matsaung, G., 2002. Assessment of the microbial quality of river water sources in rural Venda communities in South Africa. WaterSA 28 (3), 287–292.
Okawa, Y., Ohtomo, Y., Tsugawa, H., Matsuda, Y., Kobayashi, H., Tsukamoto, T., 2004. Isolation and characterization of a cytotoxin produced by Plesiomonas shigelloides P-1 strain. FEMS Microbiol. Lett. 239 (1), 125–130.

Olıas, M., Nieto, J.M., Sarmiento, A.M., Cerón, J.C., Cánovas, C.R., 2004. Seasonal water quality variations in a river affected by acid mine drainage: the Odiel River (South West Spain). Sci. Total Environ. 333 (1–3), 267–281.
Ostfeld, R.S., Keesing, F., Eviner, V.T. (Eds.), 2010. Infectious Disease Ecology: Effects of Ecosystems on Disease and of Disease on Ecosystems. Princeton University Press.
Park, Y., Cho, K.H., Park, J., Cha, S.M., Kim, J.H., 2015. Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. Sci. Total Environ. 502, 31–41.
Pratt, B., Chang, H., 2012. Effects of land cover, topography, and built structure on seasonal water quality at multiple spatial scales. J. Hazard Mater. 209, 48–58.
Prayitno, S.B., Latchford, J.W., 1995. Experimental infections of crustaceans with luminous bacteria related to *Photobacterium* and *Vibrio*. Effect of salinity and pH on infectiosity. Aquaculture 132 (1–2), 105–112.
Rahmati, O., Choubin, B., Fathabadi, A., Coulon, F., Soltani, E., Shahabi, H., Mollaefar, E., Tiefenbacher, J., Cipullo, S., Ahmad, B.B., Bui, D.T., 2019. Predicting uncertainty of machine learning models for modelling nitrate pollution of groundwater using quantile regression and UNEEC methods. Sci. Total Environ. 688, 855–866.
Ramos, M.C., Quinton, J.N., Tyrrel, S.F., 2006. Effects of cattle manure on erosion rates and runoff water pollution by faecal coliforms. J. Environ. Manag. 78 (1), 97–101.
Ren, L., Jeppesen, E., He, D., Wang, J., Liboriussen, L., Xing, P., Wu, Q.L., 2015. pH influences the importance of niche-related and neutral processes in lacustrine bacterioplankton assembly. Appl. Environ. Microbiol. 81 (9), 3104–3114.
Rippey, S.R., Cabelli, V.J., 1980. Occurrence of Aeromonas hydrophila in limnetic environments: relationship of the organism to trophic state. Microb. Ecol. 6 (1), 45–54.
Ruiz-Moreno, D., Pascual, M., Bouma, M., Dobson, A., Cash, B., 2007. Cholera seasonality in Madras (1901–1940): dual role for rainfall in endemic and epidemic regions. EcoHealth 4 (1), 52–62.
Shinohara, T., Okamoto, K., Koyano, S., Otani, A., Yamashita, M., Wakimoto, Y., Jubishi, D., Hashimoto, H., Ikeda, M., Harada, S., Okugawa, S., 2021. Plesiomonas shigelloides septic shock following ingestion of Dojo nabe (loach hotpot). In: Open Forum Infectious Diseases, vol. 8. Oxford University Press, US. No. 8, p. ofab401.
Sierralta Chichizola, V., Mayta Huatuco, E., León Quispe, J., 2016. First report of Plesiomonas shigelloides as opportunistic pathogen in tilapia Oreochromis niloticus (Linnaeus, 1758) in a fish farm in Lima, Peru. Rev. Invest. Vet. Perú 27 (3), 565–572.
Sinkko, H., Lukkari, K., Sihvonen, L.M., Sivonen, K., Leivuori, M., Rantanen, M., Paulin, L., Lyra, C., 2013. Bacteria contribute to sediment nutrient release and reflect progressed eutrophication-driven hypoxia in an organic-rich continental sea. PLoS One 8 (6), e67061.
Srinivasan, R., Lalitha, T., Brintha, N.C., Sterlin Minish, T.N., Al Obaid, S., Alharbi, S.A., Sundaram, S.R., Mahilraj, J., 2022. Predicting the Growth of F. Proliferatum and F. Culmorum and the Growth of Mycotoxin Using Machine Learning Approach. BioMed Research International, 2022.
Stock, I., 2004. *Plesiomonas shigelloides*: an emerging pathogen with unusual properties. Rev. Med. Microbiol. 15 (4), 129–139.
Sun, Y., Clarke, B., Clarke, J., Li, X., 2021. Predicting antibiotic resistance gene abundance in activated sludge using shotgun metagenomics and machine learning. Water Res. 202, 117384.
Taylor, D.N., Trofa, A.C., Sadoff, J., Chu, C., Bryla, D., Shiloach, J., Cohen, D., Ashkenazi, S., Lerman, Y., Egan, W., 1993. Synthesis, characterization, and clinical evaluation of conjugate vaccines composed of the O-specific polysaccharides of *Shigella dysenteriae* type 1, *Shigella flexneri* type 2a, and *Shigella sonnei* (*Plesiomonas shigelloides*) bound to bacterial toxoids. Infect. Immun. 61 (9), 3678–3687.
Tong, S.L., Lu, Y., 2000. Global climate change and infectious disease (in Chinese). Chin. J. Dis. Control Prev. 4, 17–19.
Toro, M., Weller, D., Ramos, R., Diaz, L., Alvarez, F.P., Reyes-Jara, A., Moreno-Switt, A.I., Meng, J., Adell, A.D., 2022. Environmental and anthropogenic factors associated with the likelihood of detecting Salmonella in agricultural watersheds. Environ. Pollut. 306, 119298.
Tran, V., Saad, T., Tesfaye, M., Walelign, S., Wordofa, M., Abera, D., Desta, K., Tsegaye, A., Ay, A., Taye, B., 2022. *Helicobacter pylori* (H. pylori) risk factor analysis and prevalence prediction: a machine learning-based approach. BMC Infect. Dis. 22 (1), 1–14.
Vezzulli, L., Colwell, R.R., Pruzzo, C., 2013. Ocean warming and spread of pathogenic *Vibrios* in the aquatic environment. Microb. Ecol. 65 (4), 817–825.
Wang, X., Xu, L., Cao, H., Wang, J., Wang, S., 2013. Identification and drug sensitivity of a *Plesiomonas shigelloides* isolated from diseased sturgeons. Wei sheng wu xue bao= Acta microbiologica Sinica 53 (7), 723–729.
Whitman, R.L., Nevers, M.B., 2003. Foreshore sand as a source of *Escherichia coli* in nearshore water of a Lake Michigan beach. Appl. Environ. Microbiol. 69 (9), 5555–5562.
Wilhelm, R.C., van Es, H.M., Buckley, D.H., 2022. Predicting measures of soil health using the microbiome and supervised machine learning. Soil Biol. Biochem. 164, 108472.
World Health Organisation, 2009. Global Health Risks: Mortality and Burden of Disease Attributed to Selected Major Risks. Available online: http://www.who.int/healthinfo/global_burden_disease/GlobalHealthRisks_report_full.pdf. (Accessed 20 April 2019).
World Health Organization, 2008. Safer water, better health: costs, benefits and sustainability of interventions to protect and promote health. WHO. https://apps.who.int/iris/bitstream/handle/10665/43840/9789241596435_eng.pdf?sequence=1.
Wu, J., Song, C., Dubinsky, E.A., Stewart, J.R., 2021. Tracking major sources of water contamination using machine learning. Front. Microbiol. 3623.
Wu, S., Feng, J., Liu, C., Wu, H., Qiu, Z., Ge, J., Sun, S., Hong, X., Li, Y., Wang, X., Yang, A., 2022. Machine learning aided construction of the quorum sensing communication network for human gut microbiota. Nat. Commun. 13 (1), 1–13.
Yang, Y., Shen, Z., Bissett, A., Viscarra Rossel, R.A., 2022. Estimating soil fungal abundance and diversity at a macroecological scale with deep learning spectrotransfer functions. Soils 8 (1), 223–235.
Ye, L., Cai, Q., Zhang, M., Tan, L., 2014. Real-time observation, early warning and forecasting phytoplankton blooms by integrating in situ automated online sondes and hybrid evolutionary algorithms. Ecol. Inf. 22, 44–51.
Yi, H.S., Lee, B., Park, S., Kwak, K.C., An, K.G., 2019. Prediction of short-term algal bloom using the M5P model-tree and extreme learning machine. Environmental Engineering Research 24 (3), 404–411.
Yin, Z., Zhang, S., Wei, Y., Wang, M., Ma, S., Yang, S., Wang, J., Yuan, C., Jiang, L., Du, Y., 2020. Horizontal gene transfer clarifies taxonomic confusion and promotes the genetic diversity and pathogenicity of *Plesiomonas shigelloides*. mSystems 5 (5) e00448-20.
Yoo, K., Yoo, H., Lee, J.M., Shukla, S.K., Park, J., 2018. Classification and regression tree approach for prediction of potential hazards of urban airborne bacteria during Asian dust events. Sci. Rep. 8 (1), 1–11.
Zhang, G., Lu, Y., 2012. Bias-corrected random forests in regression. J. Appl. Stat. 39 (1), 151–160.
Zhou, J., Liu, W., Deng, Y., Jiang, Y.H., Xue, K., He, Z., Van Nostrand, J.D., Wu, L., Yang, Y., Wang, A., 2013. Stochastic assembly leads to alternative communities with distinct functions in a bioreactor microbial community. mBio 4 (2) e00584-12.
Zhou, L., Zhao, Z., Shao, L., Fang, S., Li, T., Gan, L., Guo, C., 2022. Predicting the abundance of metal resistance genes in subtropical estuaries using amplicon sequencing and machine learning. Ecotoxicol. Environ. Saf. 241, 113844.
Zhu, J., Li, H., Jing, Z.Z., Zheng, W., Luo, Y.R., Chen, S.X., Guo, F., 2022. Robust host source tracking building on the divergent and non-stochastic assembly of gut microbiomes in wild and farmed large yellow croaker. Microbiome 10 (1), 1–15.
Zhuang, X., Zhou, S., 2019. The prediction of self-healing capacity of bacteria-based concrete using machine learning approaches. Comput. Mater. Continua (CMC) 59 (2019). *Nr. 1*.