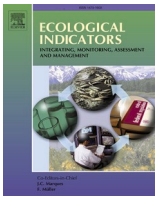
Ecological Indicators 150 (2023) 110201

Contents lists available at ScienceDirect

Ecological Indicators

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



Lotka-Volterra analysis of river Ganga pollution in India

Norbert Brunner [a,*], Sukanya Das [b], Markus Starkl [a]

[a] University of Natural Resources and Life Sciences, Vienna, Austria
[b] TERI School of Advanced Studies, New Delhi, India

ARTICLE INFO

Keywords:
Differential equation model
Ganges River
Statistical correlation
Water pollution
Water quality index

ABSTRACT

Water quality indices (WQI) are a useful tool to assess river water pollution. We defined pollution shares (indicating the relative importance of pollutants) from a WQI and studied their dynamics. Using open data from 2012 to 2020 for 105 monitoring stations along river Ganga in India, we fitted systems of generalized Lotka-Volterra (LV) differential equations to these shares. We used autonomous LV-systems (the interaction coefficients were constant) and LV-systems with variable (linear) interaction coefficients. 28 of the 105 stations had sufficient data for these models, whereby for 10 stations the autonomous system fitted well to all timeseries of the eight considered pollutants, and for 9 stations the model with linear interaction coefficients. For them we defined three candidates for "importance-growth indicators": (a) the interaction coefficients of the autonomous LV-system, (b) the leading coefficient of the interaction coefficient of the system with linear coefficients, and (c) the roots of these linear coefficients. We explored the variability of the indicators and applied them to identify stations with a similar temporal evolution of the pollutants. Further, we suggested applications to wastewater management, as at several stations the indicators (a) and (b) forecasted an increasing relative importance of nitrates/nitrites, which currently pose no problems but finally may require an upgrading of existing wastewater treatment.

1. Introduction

Starting with Horton (1965) many different water quality indices have been developed to summarize and report water quality data in a consistent manner (Bharti and Katyal, 2011). Typically, the indices have the form of weighted arithmetical or geometrical means, relating the observed pollution concentrations to certain thresholds (with a higher value of the index indicating more pollution), depending also on the available data for a particular river and the intended use of its water. There is a vast literature applying such indices to assess the changes in water quality at specific study sites during given timespans (e.g., Beck et al., 2019; Chen et al., 2020; Hasan et al., 2020; Parween et al., 2022).

Other than the above-mentioned literature, we focus on the relative importance of pollutants and possible implications for water management. Thereby, we measure the relative importance of a pollutant by its "pollution share", meaning its share of the relevant water quality index (here: index of Section 2.2 below). The growth of the relative importance of a pollutant is an early indicator of a future pollution risk, even if currently this pollutant is still negligible. Detecting such risks early allows for a timely planning of management actions, such as an upscaling of water treatment infrastructure.

The systematic search for pollutants with emerging relative importance is a recent topic of water management, starting with Tang et al. (2022), who studied past deterioration rates of water quality parameters at a given site. Here, we develop a different approach and define "importance-growth indicators" with the intention to summarize the long-term evolution of the relative importance of all considered pollutants at a given site and to identify the pollutant with the highest future risk. This task of identifying the direction of trends differs fundamentally from forecasting pollution levels, where in some cases even reliable forecasts over three days may be difficult to obtain (e.g., Shamshirband et al., 2019). To define the new indicators, we propose a dynamical model, that describes the temporal evolution of the pollution shares of all considered pollutants at a certain site simultaneously by means of a system of deterministic (generalized) Lotka-Volterra (LV) differential equations. Thereby, for each except one pollutant (the

Abbreviations: β, Blomqvist beta (correlation); BOD, bio-chemical oxygen demand; Con, electrical conductivity; DO, dissolved oxygen; FC, fecal coliforms; LV, Lotka-Volterra; pH, potential of hydrogen; *PR*, and *PR**, precision rate; R^2, R-squared; *SSE*, sum of squared errors; TC, total coliforms; UP, Uttar Pradesh; WB, West Bengal; *WQI*, water quality index.

* Corresponding author.
E-mail address: norbert.brunner@boku.ac.at (N. Brunner).

https://doi.org/10.1016/j.ecolind.2023.110201
Received 12 December 2022; Received in revised form 23 March 2023; Accepted 29 March 2023
Available online 5 April 2023

"outside pollutant", whose shares are the remainder of the others from 100 %) the evolution of its pollution share is described by another differential equation of the system. The "importance-growth indicators" are then defined from the best-fit parameters of these differential equations. Thus, we are combining several aggregations: The pollution data define a water quality index used to define the pollution shares of each pollutant at that year and site. Fitting the LV-model to these timeseries then defines the "importance-growth indicators" of the pollutants that may be used to forecast the future relative importance of each pollutant at the given site. The paper explores several candidate definitions of such indicators. The indicators are highly aggregating, transforming information from eight timeseries at a station into seven numbers.

LV-models are a common tool in mathematical ecology, including for the study of aquatic ecosystems. (Example from Huang et al., 2020: How did water pollution by microparticles alter the aquatic predator–prey dynamics?) Later, they were adopted for business applications: Starting with Modis (1999), LV-systems were used to characterize the dynamics of stock market shares. Our approach differs from both approaches with respect to the interpretation of the models: The drivers of the pollution dynamics are neither purely ecological (as for the classical predator–prey models of animal populations) nor purely social (as for stock market shares). Rather, they result from a combination of ecological, socio-economic, and institutional influencing factors. Of course, there are many other classes of mathematical models that might be adopted to forecast pollution shares or identify trends for them (e.g., time series analysis using moving averages or exponential smoothing, simple regression, logit regression, other more complex trend models, such as a model for market diffusion by Bass, 1969, or models from systems dynamics; see Sterman, 2001). We decided to use LV-models, as reviews have shown that such models produced reasonable forecasts from small datasets (Dang et al., 2016; Fu et al., 2017). Moreover, we used a clever variant of the LV-equations from Marasco et al. (2016). Although this variant added flexibility to the model by allowing for variable coefficients, it led to considerable simplifications in practical applications, as its system of differential equations could be solved analytically. Further, the fitting of the model parameters to given empirical data (a difficulty with the conventional LV-model, c.f. Kloppers and Greeff, 2013) could be done in a straightforward way resembling logit regression. This resulted in multiple applications, such as in grocery business (Horgan, 2020), automotive industry (Ziegler et al., 2020), food industry (Bauer et al., 2022), or water management (Focacci and Quintavalla, 2020). The first author of this paper has used this model repeatedly, too, and some more technical comments (Section 2.4) draw on these experiences. However, as the interpretation of pollution shares is fundamentally different from the interpretation of market shares that was hitherto used, the application of this model to define importance-growth indicators is new.

To illustrate our approach in a concrete setting, using data from the public domain, we studied water pollution of river Ganga (Ganges River). In India, this is an important policy issue, as the river provides water for 43 % of India's population (GoI, 2009). Thus, there is a large body of literature about this topic (Google Scholar: 20,000 papers since 2013). The Government of India (CPCB, 2023) provides open data about water pollution of all major rivers. For river Ganga, these data are annual summaries of the water quality monitoring at 105 stations during the years 2012 to 2020. However, there were specific issues with the data. First, the latest data were from 2020, whence the impact of recent infrastructure investments could not be assessed. Second, the timeseries was quite short for a study of the temporal evolution of pollution; at most nine data points at each monitoring station (one for each year). To response to this challenge, we removed 2/3 of the stations with unsuitable (e.g., incomplete) data from our study. Fig. 1 plots the resulting study site; it focused on Uttar Pradesh (UP) and West Bengal (WB). Third, the data delivered only values of certain "core parameters" for water quality.

In view of these issues, we defined the goal of our paper as follows: Its purpose is a proof of principle, namely that the importance-growth indicators provide meaningful information, even if the data are not perfect.

Fig. 1. Study area after choosing suitable stations; highlighted areas: Uttar Pradesh and West Bengal. The capital city Delhi is shown for better orientation. Plot using Mathematica 13.2, based on Open Street Map.

2. Method

2.1. Data

The webpage of the Central Pollution Control Board, CPCB (2023) informs about river Ganga water quality at 105 monitoring stations, identified by a code number, a rough description of its location, and the state. The code numbers in general increase with the flow-direction of the river, as it passes through five Indian states, Uttarakhand (UK), Uttar Pradesh (UP), Bihar (BR), Jharkhand (JH), and West Bengal (WB), thereby traveling 2304 km (Mirza, 2004) from the source of river Bhagirathi at Gaumukh, UK, to the Bangladesh border. At Farakka, WB, 10 km before the border to Bangladesh, a barrage diverts about half of the water to river Huglī (Hooghly River). The data from WB inform about this final Indian branch (260 km) of river Ganga flowing into the Bay of Bengal. Further, there are also stations at the main headwaters, Alaknanda and Bhagirathi rivers. (At their confluence in Devprayag, UK, the river acquires the name Ganga.).

The data collect annual minimal and maximal values (for 2012 to 2014 also average values) of certain indicators from 2012 to 2020. (More recent river water quality data for 2021 were not yet available for download; we checked this on 01.03.2023. Data for 2015 linked to a missing page. We retrieved them from a PDF file at the water quality management site of the webpage.) The indicators are water temperature (abbreviation Temp, °C), dissolved oxygen (DO, mg/L, L = liter), potential of hydrogen (pH), conductivity (Con, μS/cm, S = Siemens), biochemical oxygen demand (BOD, mg/L), nitrate/nitrite (N, mg/L), fecal and total coliforms (FC and TC in MPN/100 mL, MPN = most probable number); see Section 6 for a discussion. The data of 2019 and 2020 informed also about fecal streptococci. As their time-series was too short for an analysis of the temporal evolution, we disregarded streptococci.

In total, the data from 105 measurement stations defined a table with 652 lines for the pollution at a station and year with data for 29 to 99 stations per year. From this table we removed the columns with information that was irrelevant for our water quality index (e.g., annual minimum of BOD). However, there were gaps in the data. For example, for 2012, our source provided only data between Gangotri, UK, and Varanasi, UP; the data for the other three states apparently were lost. As we wished to compare pollution since 2012 in a coherent manner, we first removed all lines from the original table, where one of the relevant parameter values was missing. Second, we removed all lines, where an index-value was zero. For, such zeroes could be gaps (auto-filing empty cells with 0 s) and they did not allow logit-style transformations, which we later used. Third, as the study of trends would be futile if based on only few points of time, from the remaining table we removed all stations with data for four or fewer years. The Supporting Data collects the retained data and informs about which data were removed, and why (File SD1).

This resulted in a smaller table with data from 28 stations (215 lines). Fig. 1 plots the resulting study area. These were 17 stations in UP (1062 at Garhmukteshwar, 2488 and 2489 at Anoopshahar/Anupshahr, 1145 at Narora, 2490 at Kachhla, 1063 and 1066 at Kannauj, 1146 at Bithur/Bithoor, 1067 and 1068 at Kanpur/Cawnpore, 1147 at Dalmau, 2498 at Kalakankar near Raebareli, 1046, 2487, and 1049 at Allahabad/Prayagraj, 2485 and 2486 at Mirzapur, and 1071 at Varanasi/Banaras), and 11 stations in WB (1080 at Berhampore/Baharampur, 2506 at Tribeni, 1054, 1472, 1053, 2511, 1471, and 1470 at Kolkata/Calcutta, 1052 at Uluberia, and 1469 at Diamond Harbor/Diamond Harbour).

2.2. Water quality index

As was mentioned in the introduction, there are multiple water quality indices, with one even specifically designed for river Ganga (Ved Prakash index: Abbasi and Abbasi, 2012, Bhutiani et al., 2016). However, the goal of our paper is a proof of principle regarding indicators for relative-importance growth. Therefore, for us it did not matter, which water quality index was used, if it was reasonable. Consequently, we defined a water quality index ad hoc with the intention to utilize the available data. (This is not the Ved Prakesh index.).

We first defined indices that aimed at transforming the measurements into numbers that were comparable with respect to the potential damages. We defined them so that index value 0 was optimal, while an index value 1 or higher was inacceptable, meaning a pollution (concentration) above a legal or a common non-legal threshold. Such thresholds may derive from standards that aim at ensuring that certain ecological services (e.g., bathing, fishery) are safe (Starkl et al., 2013; Starkl et al., 2018a; Cid et al, 2022).

- Temperature (Temp) may modify the impact of other pollutants. As 40 °C (a non-legal threshold) is not in the temperature range preferred by fish, while 20 °C or below is OK, we defined an index $WQ_0(x_0) = x_0/20 - 1$. The relevant parameter from our data was x_0: annual maximum of the observed temperatures.
- Fecal coliforms (FC) are a health challenge. CPCB (2023) refers to a legal threshold of at most 2500 MPN/100 mL in India for fecal coliforms in bathing water. We defined $WQ_1(x_1) = x_1/2500$. The relevant parameter from our data was x_1: annual maximum of fecal coliform concentration.
- For total coliforms (TC), CPCB (2023) reported a threshold of at most 5000 MPN/100 mL. We defined $WQ_2(x_2) = x_2/5000$. The relevant parameter from our data was x_2: annual maximum of total coliform concentration.
- Nitrate/nitrite (N) is a toxicant for fish and associated to algal bloom. A common (non-legal) threshold is 30 mg/L. We defined an index $WQ_3(x_3) = x_3/30$. The relevant parameter from our data was x_3: annual maximum of *N*-concentration.
- Biochemical oxygen demand (BOD) measures organic pollution. CPCB (2023) refers to a legal threshold of at most 3 mg/L for bathing (up to 12 mg/L would be conceivable for fishery). We defined an index $WQ_4(x_4) = x_4/3$. The relevant parameter from our data was x_4: annual maximum of BOD concentration.
- Conductivity (Con) informs about salinity. Based on a (non-legal) threshold for river water, we defined $WQ_5(x_5) = x_5/4000$. The relevant parameter from our data was x_5: annual maximum of conductivity.
- Dissolved oxygen is necessary for aquatic life. CPCB (2023) refers to a legal threshold of at least 5 mg/L for bathing. As values above 10 mg/L are optimal for fishery, we defined $WQ_6(x_6) = 2 - x_6/5$. The relevant parameter from our data was x_6: annual minimum of the concentration of dissolved oxygen.
- The potential of hydrogen (pH) informs about the acidity or basicity. CCPC(2022) refers to the legally acceptable range of pH between 6.5 and 8.5 (optimal for fish: pH = 8.2). We defined $WQ_7(x_7, x_8) = \max\{(82 - 10 \times x_7)/17, (10 \times x_8 - 82)/3\}$. The relevant parameters from our data were x_7 and x_8: annual minimum and annual maximum of the pH values (Fig. 2 explains the index).

The sum of these indices defined our ad hoc index to assess total pollution. Thus, for each station and year, the input was a vector, $\boldsymbol{x} = (x_0, x_1, \ldots, x_8)$; a line in the table of SD1. We used it to define our water quality index *WQI* as in equation (1).

$$WQI(x) = WQ_0(x_0) + WQ_1(x_1) + \ldots + WQ_7(x_7, x_8) \quad (1)$$

For the selected 28 stations (Section 2.1), the values of these indices are provided as Supporting Data (in file SD1). To measure the relative importance of each pollutant, we defined its "pollution share", s_i, as the relative contribution to total pollution: $s_i = WQ_i/WQI$. The values of these pollution shares were $s_i(t_k)$ for pollutants $i = 0, 1, \ldots, 7$ at years t_k chosen from 2012, 2013, …, 2020 (at different stations data were available for different sequences of years). They are tabulated in the Supporting Data, too (file SD1). The temporal evolution of these shares

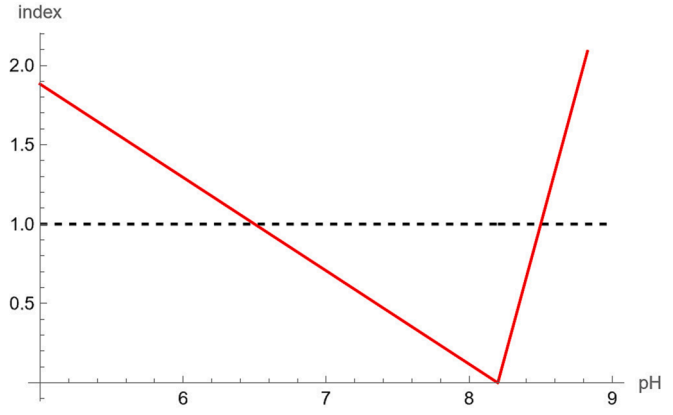


Fig. 2. Idea behind the formula for the definition of index WQ_8 for the pH-value (red). The dashed line indicates the index value at the legal thresholds. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

is the object of our study.

2.3. Lotka-Volterra model

Marasco et al. (2016) proposed the following variant (2) of the LV-system of differential equations:

$$\frac{dy_i(t)}{dt} = y_i(t) \bullet \left(g_i(t) - \sum_{j\geq 1} g_j(t) \bullet y_j(t) \right) i = 1, 2, \cdots \tag{2}$$

We used the solutions, $y_i(t)$, of this system of differential equations to describe the temporal evolution of pollution shares ("inside pollutants": $i = 1, 2, \ldots$) over time, t, at a given observation station. To this end, we first defined an "outside pollutant" (the others being "inside pollutants"). We used temperature for this purpose (renumbering: index $i = 0$), as discussed further in Section 4.

We then defined (dis)utilities relative to temperature by means of equation (3), where $s_i = WQ_i/WQI$ was the pollution share for pollutant, i. By this definition, the (dis)utility of the outside good was set to be zero. (In models, where only two pollutants are considered, $i = 0$ and $i = 1$, the utility $u_1(t)$ is the logit of s_1.) The Supporting Data tabulates the (dis) utilities, $u_i(t_k)$, whereby $i = 1, 2, \ldots, 7$ lists the inside pollutants and t_k are the years (from 2012 to 2020) considered for the respective stations (file SD1).

$$u_i(t) = \ln\left(\frac{s_i(t)}{s_0(t)}\right) i = 1, 2, \cdots \tag{3}$$

The parameters of model (2) will be determined from fitting suitable curves to the time-series of the observed (dis)utilities (3) whereby we could handle the pollutants independently from each other. We fitted simple functions, $v_i(t)$, to the observed (dis)utilities (nonlinear regression). For reasons explained in Section 4, we used only linear or quadratic polynomials to model (dis)utilities. (The model curves, y_i, were not polynomials.).

We then defined "interaction coefficients", $g_i(t)$, as the derivatives of these polynomials, equation (4). For linear polynomials, using these interaction coefficients in equation (2) defined an autonomous system with constant coefficients.

$$g_i(t) = \frac{dv_i(t)}{dt} i = 1, 2, \cdots \tag{4}$$

The following functions, $y_i(t)$ of equation (5), were then solutions of the LV-system (2) and they approximated the pollution shares. The pollution shares of the outside pollutant, $y_0(t)$, were approximated by the remainder of these shares.

$$y_i(t) = \frac{\exp(v_i(t))}{1 + \sum_{j\geq 1} \exp(v_j(t))} i = 1, 2, \cdots \tag{5}$$

2.4. Implementation and statistical methods

Our computations and plots used Mathematica 13.2 of Wolfram Research (2023). For data fitting, we used its function NonlinearModelFit (for linear problems it automatically used linear programing). It applied the method of least squares to fit regression polynomials (and other types of functions) to the observed (dis)utilities, u_i. Thereby, for a given station and for each good, i, we sought parameters for the regression polynomial, $v_i(t)$, that minimized the sum of squared errors, SSE_i of equation (6). Here, t_k (for $k = 1, 2, \ldots, K$) were the years of measurement (which were not the same for all stations). We reported goodness of fit using R-squared, equation (7).

$$SSE_i = \sum_{k=1}^{K} (v_i(t_k) - u_i(t_k))^2 \tag{6}$$

$$R_i^2 = 1 - \frac{SSE_i}{\sum_{k=1}^{K} (u_i(t_k) - \text{mean}(u_i(t_1), u_i(t_2), \cdots))^2} \tag{7}$$

As is well-known, water quality data in general are subject to errors and uncertainty (Jha and Jha, 2013), resulting in outliers. Maximal and minimal values are particularly sensitive to irregularities. Some authors suggest AI methods to "correct" the data (Kouadri et al., 2021). Instead, we used Mathematica to identify and remove potential outliers to improve the fit: For each station and a given pollutant, we identified the datapoint with the least single deletion variance (provided by the "NonlinearModelFit" command). Owing to the small number of available datapoints, we added a test, if we should retain this datapoint: We compared it with the median of the single deletion variances. (We conducted a location test, whereby Mathematica selected a test statistic automatically from several options, such as a sign test. If it resulted in a P-value below 0.01, we deemed the year with this datapoint as strange. Note, that this was not the P-value of the applied test, which varied with the sample size: The smaller the sample, the higher the chances were that the suspicious datapoint would pass the test.) We then removed all data for this year (also for the other pollutants) from the station record. If there remained less than 3 (linear model) or 4 (quadratic model) years, then we refrained from using the given model for that station.

For the quadratic regression polynomials, we were also interested in their vertices (roots of the interaction coefficients, i.e., of the first derivatives of the regression polynomials). However, to compute their confidence intervals, we had to re-parametrize the polynomials, equation (8), using instead of parameters p_0, p_1, and p_2 (equation (8), left) the new parameters $a = p_2$, $b = p_0$ and vertex $v = -0.5 \cdot p_1/p_2$, and solve the nonlinear regression problem in a, b, and v.

$$p_0 + p_1 \bullet t + p_2 \bullet t^2 = a \bullet t^2 - 2 \bullet a \bullet v \bullet t + b \tag{8}$$

To report the goodness of the fit of the LV-model functions, y_i of equation (5), to the observed pollution shares, s_i, we used the precision rate of equation (9): It is 1 minus the mean absolute relative error relative to the data. The precision rate was suggested by Dang et al. (2016). We also considered the precision rate PR^* relative to the model; write y_i instead of s_i in the denominators of equation (9). We applied it for the outside good, too. We did not report R-squared for the model functions (and neither the mean squared error), as they were not least-square fits to the pollution shares.

$$PR_i = 1 - \frac{1}{K} \bullet \sum_{k=1}^{K} \frac{|s_i(t_k) - y_i(t_k)|}{s_i(t_k)} \tag{9}$$

The least-squares method provided maximum-likelihood estimates, if the fit errors were "white noise", independent and identically normally distributed random variables (Diebold, 2007). We tested these

assumptions using the Anderson-Darling test for normal distributions and the Box and Pierce test for significant autocorrelations; these tests are implemented in Mathematica. (The selection of these methods took the small sample sizes into consideration. For instance, the Cramér and von Mises test for normal distributions requires 7 or more data points.).

To explore the variability of the best fit parameters, we analyzed their confidence intervals (provided by NonlinearModelFit). However, if one of the above tests for the "white noise" assumption failed, this could indicate a misspecification of the model. In this case the variability of parameters would be larger (Newey and West, 1987: broader confidence intervals). If the "white noise" tests were fulfilled, we used the asymptotic multinormal distribution of the parameters for simulations. (Its mean values and covariance matrix were provided by NonlinearModelFit.).

Consequently, prior to a more detailed analysis of our outcomes we identified the stations, where the fit residuals of the regression polynomials for all seven timeseries of (dis)utilities inside pollutants satisfied the "white noise" assumption. Moreover, for such analysis we considered only stations, where the model functions had an acceptable fit to all eight timeseries of the pollution shares. Our selection criteria were a P-value below 0.01 for the normal distribution test and the autocorrelation test, and precision rates (*PR* and *PR**) above 0.5 (see Section 4 for a discussion). Further, for a low R-squared we checked the regression polynomials. However, in general this indicated an almost constant pollution share.

There are multiple measures for the correlations between data vectors. Generally, one uses the Pearson correlation coefficient. However, in some situations there may be problems with the scaling: On face value, data may appear as ratio-scaled real numbers, but this may be misleading. Rather, correlation needs to be redefined to capture the information that is needed and available from the data. For example, Spearman rank correlation is used where only the order relations between the data matter (ordinal data). Another example from water quality assessment is "grey relational analysis" (Gai and Guo, 2023), where the compared vectors are transformed to detect geometrical similarities. In our study, when comparing different vectors of importance-growth indicators it occurred to us that for them the positions relative to the median might matter. To focus on this aspect, which did not require the full ordinal information about the vectors, we applied Blomqvist (1950) beta (β): Given vectors $\boldsymbol{x} = (x_1, x_2, \ldots)$ and $\boldsymbol{y} = (y_1, y_2, \ldots)$, it considers the vectors of signs of the differences to the respective medians, (x_i–median($\boldsymbol{x}$)) and (y_i–median($\boldsymbol{y}$)); Blomqvist β is then the (usual) Pearson correlation coefficient between these vectors of signs.

3. Results

3.1. Conclusions from WQI for water management

The large population along river Ganga caused massive water pollution from untreated domestic sewage and industrial wastewater (GoI, 2009). For, there are major urban centers along river Ganga, such as Haridwar in UK, Kannauj, Kanpur, Allahabad, and Varanasi in UP, Patna in BR, and Kolkata in WB. Further, there is pollution from industrial clusters at Kashipur and Moradabad discharging into river Ramganga, and clusters at Meerut and Modinagar discharging into river Kali. (Both rivers join Ganga near Kannauj, UP.) Till the late 1970 s there was barely an awareness for water pollution in India (Wohl, 2012). This indifference was surprising, considering the high spiritual value of the "purity" of river Ganga water (Alley, 2019; Eck, 1999). In 1985 the Ganga Action Plan was initiated, followed in 1993 by the Second Ganga Action Plan and in 1995 the National River Conservation Plan. Under these policies, infrastructure for the collection and treatment of sewage was funded, open defecation was reduced by the provision of toilets and public awareness raising, and riverfronts were developed (Dutta et al., 2020). However, an analysis of the water sector reform policies between 2002 and 2006 has found a preference for non-selective "soft" supporting measures for the implementation of better infrastructure, such as capacity building and guidance of local communities, while selective "hard" control, such as competitive bidding and monitoring to link funding to success was largely neglected, resulting in suboptimal outcomes (Brunner et al., 2010; Starkl et al., 2013). Thus, in 2015 sewage treatment capacities along river Ganga still ranged from inadequate (6.6 % to 8.9 % in BR, JH, and WB) to low (30.1 % to 37.2 % in UK and UP), whence in 2014 the (ongoing) Clean Ganga Program (Namami Gange) was initiated as a game changer (Breitenmoser et al., 2022). Supporting actions were the Clean India Mission (Swachh Bharat) of 2014, the Atal Mission for Rejuvenation and Urban Transformation (AMRUT, formerly JNNURM) of 2015, and the Smart City initiative of 2017. As these experiences with water sector reform policies have shown, it took three decades till deficiencies of past policies were recognized and addressed by new policy initiatives.

Did these policies significantly affect river water quality? We asked this question for the previously selected 28 stations with sufficient data (Section 2.1). To identify possible trends, we fitted linear regression lines to the timeseries of our ad hoc defined *WQI* (Section 2.2). For 23 (82 % of the 28 stations) the regression line was decreasing (increasing for the other five stations), meaning a trend for improving water quality. However, the sign of the slope was significant for only five regression lines. Of them there were four significant improvements (i.e., 14 % of 28 stations) with a negative sign (negative upper 95 % confidence limits for the slopes at stations 1052, 1472, 1469, and 2490) and one with a positive sign (positive lower 95 % confidence limit for the slope at station 2506). Fig. 3 plots two of the regression lines with significant slopes.

Thus, when considering only those pollutants that CPCB (2023) deemed as noteworthy to report, then in general during 2012 to 2020 river Ganga pollution did not change much: In general, it did not deteriorate significantly, and there were merely 14 % significant improvements. This observation was also supported by the in general small R-squared values for the regression lines ($R^2 \leq 0.2$ at 12 stations), suggesting close to constant water quality levels. Further, a detailed look at the Supporting Data table of *WQI*-values (file SD1) suggests several issues. For 93 % (198) of its 214 data lines, at least one of the eight water quality parameters did not satisfy the water quality thresholds mentioned in Section 2.2. Most failures were observed for BOD, followed by FC and TC (193, 163 and 157 failures). For FC and TC, the failures were most dramatic, with 13–17 % of the data lines displaying coliform counts that exceeded the respective legal threshold by the factor of 100. DO and pH were problematic, too (72 and 76 failures), while N, Temp, and Con, were in general unproblematic (no, one, and nine failures, respectively).

3.2. Best fitting models and stations with acceptably fitting models

In the sequel, by the linear or quadratic LV-model we mean the LV-model, where linear or quadratic polynomials, respectively, were fitted to the (dis)utilities. The resulting solutions of the LV differential equations, functions $y_i(t)$ for the pollution shares, were not linear or quadratic. A table in the Supporting Data lists the outcomes of our computations (file SD2). For each of the 28 selected stations (Section 2.1) and each of the eight considered pollutants (Section 2.2), it lists the model parameters together with their 95 % confidence intervals (computed from data file SD1), P-values from testing the fit residuals of the regression polynomials, and precision rates of the LV-model functions. For the identification of the stations, we reported the village names and the station codes and added new station numbers in the direction of the river flow. For further conclusions we highlighted 16 of the 28 stations, where the LV-models had acceptable fits to the data for all eight model curves (we employed several tests to assess the fit).

We first fitted linear or quadratic polynomials to the (dis)utilities of (inside) pollutants and searched for outliers. Specifically for each station, we identified "strange years", where one of the seven considered

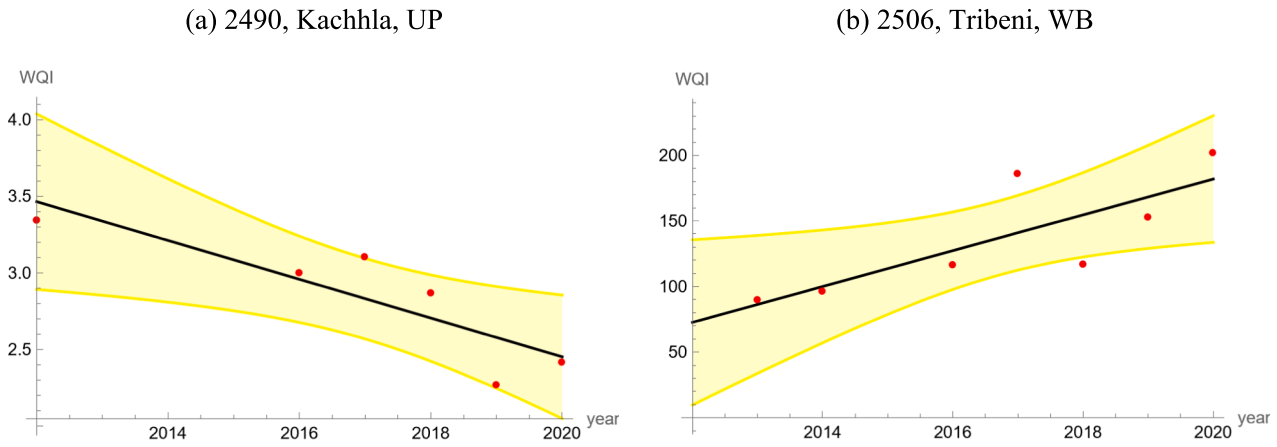


Fig. 3. Plot of the timeseries of the index values of *WQI* (red dots), the regression line (black), and 95% mean prediction bands (yellow) for two monitoring stations. Plot using Mathematica 13.2, using SD1 as data source. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

time-series possibly had an outlier (Supporting Data, file SD1). Thereby, given the regression polynomial, we identified the datapoint with the least single deletion variance. We then used a location test to check, if that datapoint might be an outlier. If so, the year, where this datapoint was observed, was labeled as "strange" and all data for strange years were removed for the station at hand. Sometimes we removed potential outliers whose deviations from the regression polynomials were within the range of random fluctuations; this could nevertheless improve the fit (Fig. 4).

The results of this step (together with the best-fit parameters of the regression polynomials) are summarized in the Supporting Data (file SD2). We found up to four strange years per station. Next, for each station we removed all data of the strange years. Finally, we removed all stations with data for less than four (linear LV-model) or five (quadratic LV-model) remaining years. (For example, for station 1063 in Kannauj, UP, the four years 2012, 2013, 2014, and 2016 were identified as strange for the linear LV-model, because there were potential outliers in 2012 for FC, in 2013 for pH, in 2014 for N, and in 2016 for DO. There were only data for two years remaining: 2015 and 2016. Consequently, for the linear LV-model station 1063 was removed for lack of data. Similarly, for the quadratic models the stations 1053 and 1472, both in Kolkata, WB, were removed.) The Supporting Data informs about the finally retained data (file SD1 with columns for the linear and quadratic LV-models, respectively).

Second, we studied the remaining data and fitted linear (and quadratic) polynomials to the (dis)utilities (c.f., black dots and black line in Fig. 4). The outcomes are listed in a table of the Supporting Data (file SD2). Thereby, for each station and each of the seven pollutants, we computed the best-fit parameters of the regression polynomials, the 95 % confidence intervals, R-squared, and the P-values for the statistical tests (Section 2.4) if the fit residuals were normally distributed and not autocorrelated. With respect to these tests, the linear LV-model and the quadratic LV-model were problematic (one test with P-value p less than 0.01) at up to two stations (misspecification of the model for at least one of the seven time-series per station). Third, using equation (5) and the regression polynomials, $v_i(t)$, we approximated the pollution shares, $s_i(t)$, by the LV-model functions, $y_i(t)$. Next, considering the outside pollutant (Temp), too, we assessed the goodness of fit by means of the precision rates, *PR* and *PR**, computed the limits of the (modeled) pollution shares for infinite time, and collected the results in the Supporting Data (file SD2). The LV-model had an acceptable the fit, if for all eight model curves for the pollution shares the precision rates *PR* or *PR** were above 0.5.

For an illustration, Fig. 5 plots the data and models curves for station 1469 at Diamond Harbor, WB, where river Ganga flows into the Bay of Bengal. Both LV-models had acceptable fits. Comparing the precision rates (minimum of *PR* and *PR**), then the quadratic LV-model had a better fit for 5 pollutants and the linear LV-model had a better fit for the

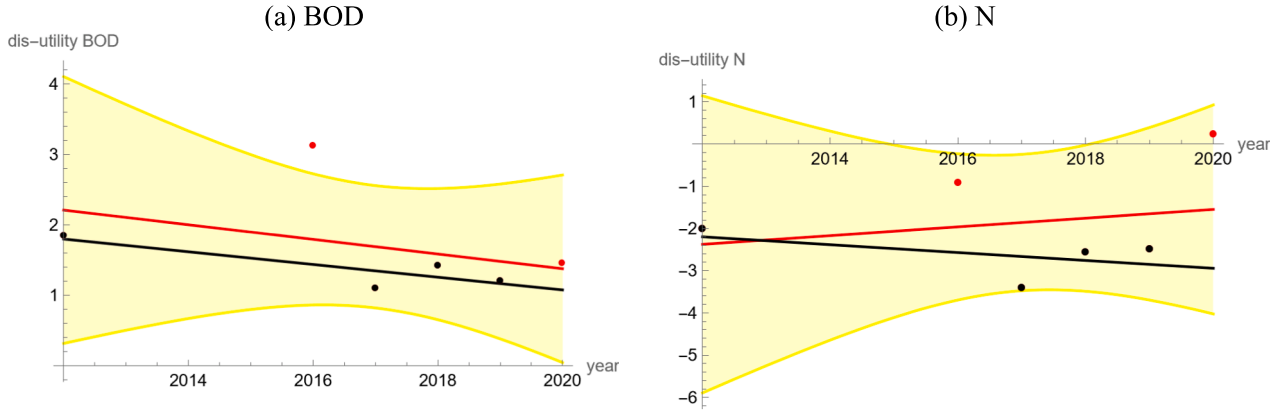


Fig. 4. Plot of the (dis)utilities for BOD and N at station 2490 (Kachhla, UP): Data for 2013–2015 had been removed due to data gaps. Red dots indicate data for strange years that were finally removed, too. Black dots are the retained data. The black and red lines are the regression lines fitted to the retained (black) and all (red and black) data, respectively. The yellow area is the 95% mean prediction band for the red line. The plots explain, why a) 2016 was strange for BOD and b) 2020 was strange for N, although in view of the mean prediction band the deviation of the latter datapoint from the red line was not excessive. Plot using Mathematica 13.2, using SD1 as data source (dots). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

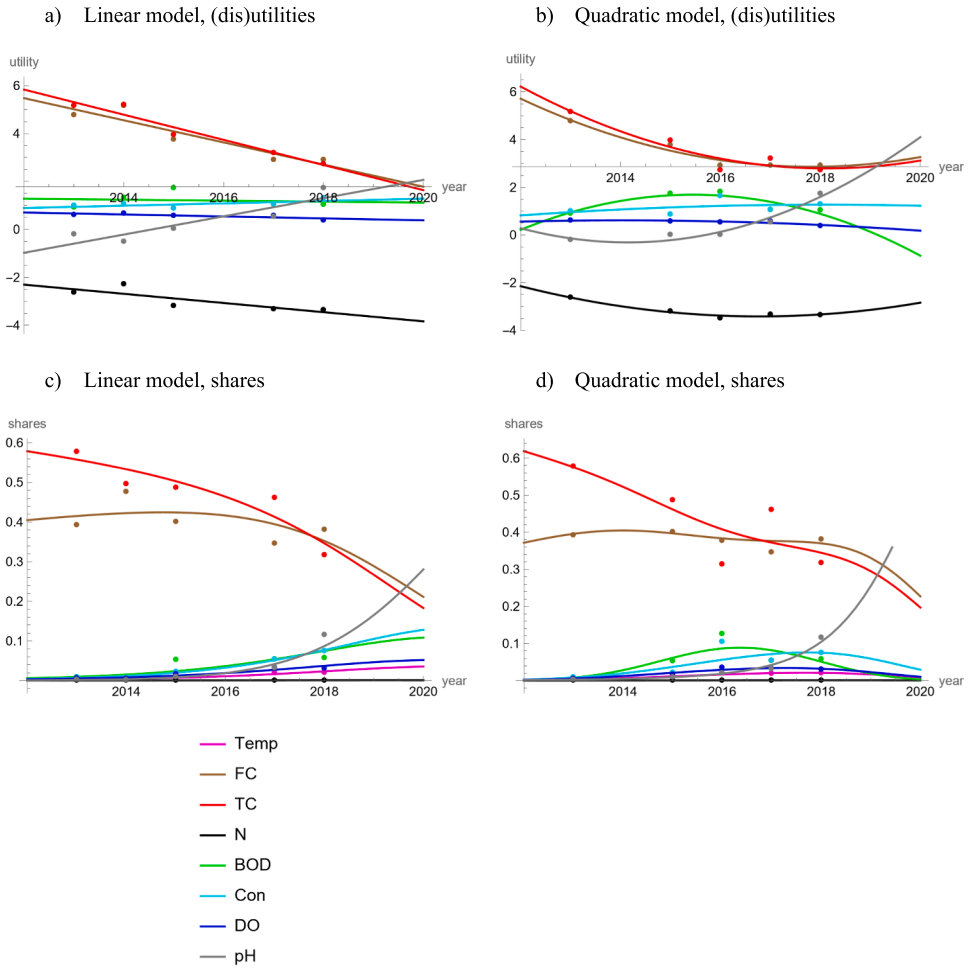


Fig. 5. Plots of the (dis)utilities, pollution shares and LV-model curves for eight pollutants at station 1469 (Diamond Harbor, WB), using the linear and quadratic LV-model functions. Outliers were removed from the data. The colors are explained in the legend (left).

other three (Con, N, TC). Thereby, both models were fitted to five years, but the sequence of years differed, as the linear and quadratic models identified different potential outliers.

We identified 10 (36 %) from 28 originally considered stations, where the linear LV-model had an acceptable fit to the data after the removal of possible outliers (strange years), and 9 (32 %), where the quadratic LV-model had an acceptable fit. For three stations both models were acceptable, for seven stations only the linear model was acceptable, and for six stations, only the quadratic model. Thus, for 16 (57 %) of the considered 28 stations one of our models had an acceptable fit. For both models, the best fit was achieved at station 2489 (Anoopshahar, UP). There, for all timeseries the P-values for the tests of normality and autocorrelations were above 0.05, the precision rates were above 0.79, and R-squared was above 0.56 (with a slightly better fit for the quadratic LV-model).

3.3. New indicators for the pollution dynamics at each station

There are two obvious candidates for importance-growth indicators: the leading coefficients of (a) the linear, and (b) the quadratic polynomials fitted to the (dis)utilities of each pollutant. Note that for the linear model the leading coefficients coincide with the interaction coefficients of the LV-system: They are the first derivatives of the fitted trend lines $v(t) = p_0 + p_1 \cdot t$. Similarly, for the quadratic model the interaction coefficients are linear functions, $g(t) = p_1 + 2 \cdot p_2 \cdot t$ (using p_0, p_1, and p_2 as the parameters of the regression polynomial). Thus, using its leading coefficient, $2 \cdot p_2$, would be equivalent to indicator (b), using p_2. Further (quadratic model), in economic literature, the roots of the interaction coefficients are of interest, as they indicate possible breaks in the market dynamics (Dominioni et al., 2019) and (dis)utility for pollutants is extremal at these roots (vertices of the parabolas of disutilities). We therefore consider (c) the roots of the interaction coefficients as another candidate for an indicator set. We refer to the indicators (a), (b) and (c) as linear, quadratic, and root indicators, respectively. We first studied them for the 16 stations, for which one of our models had acceptable fits. To explore, if a good fit was really needed, and to apply statistical tests, we also explored the indicators for the 28 stations with sufficient data.

At first, we speculated, which pollutant might dominate in the end. This means that its LV-model curve predicts a pollution share of 100 % in the limit at infinite time. For the first two indicators there is an obvious relation to the limits of the pollution shares at infinite time. As can be seen from a typical formula for a model function in Fig. 6, for the linear model the limit for $t \to \infty$ depends on the relative sizes of the leading coefficients (slopes p_1) of the linear functions in the exponents. If all slopes are negative (as in Fig. 6), the limits for the inside pollutants are 0 and so the limit for the outside pollutant is 1. If some slopes are positive, then the limit of the pollution shares is 1 for the inside pollutant with the largest slope and 0 for all other pollutants. (We disregard maxima with slope 0 or several equal maxima, as this occurs with probability 0, referring to the asymptotic multinormal distributions of the best-fit parameters.) Similarly, for the pollution shares modeled by the quadratic LV-model, the leading coefficients (p_2) of the quadratic polynomials fitted to the (dis)utilities decide about the forecasted limit,

$$\frac{e^{666.366-0.330643\,t}}{1+e^{692.438-0.343902\,t}+e^{666.366-0.330643\,t}+e^{602.03-0.298811\,t}+e^{494.977-0.246673\,t}+e^{438.476-0.216564\,t}+e^{357.056-0.176558\,t}+e^{246.601-0.121763\,t}}$$

Fig. 6. Typical model function for the pollution shares for the linear LV-model.

which (with probability 1) is 1 for one pollutant and 0, otherwise. Considering both the best fitting linear or quadratic LV-models, pH dominated 8 times, followed by N, TC, Temp (3), Con, FC (1), and BOD (0). At one station (2498) the forecasts of the linear and quadratic LV-models differed (Table 1).

Next, we compared the variability of the leading coefficients. For the 19 LV-models with acceptable fits (16 stations), the fit errors were normally distributed and independent. Assuming that these errors were accidental, then another random perturbation of the data might have led to different estimations. We modeled the best-fit parameters by random variables following the asymptotic multinormal distribution of the regression parameters. Using this distribution, we conducted 10,000 simulations of the parameters for the inside pollutants and counted, how often the leading parameter of a pollutant was largest and positive (this pollutant dominated) and how often all simulated leading parameters were negative (Temp dominated). In general (see Table 1) the dominating pollutant for the best fit model was more often dominating for the simulations than any other pollutant (exception: station 1068). However, for seven entries of the table, in at least 50 % of the simulations there were various other pollutants that by chance could be dominating, too. Consequently, also for the models with acceptable fits there remained some uncertainty about which pollutants might finally dominate.

For the 28 stations with sufficient data for our models, we assessed the (minimum) variability by the average length of the confidence intervals of the leading coefficients (upper bound minus lower bound) for the seven inside pollutants at each station. (As amongst the 28 stations we allowed for misspecifications of the model, the actual variability might have been larger.) For p_1 of the linear LV-model and p_2 of the quadratic LV-model, the logarithms of these average values were normally distributed, whereby the 95 %-quantiles for the average lengths were 1.09 and 0.77 for p_1 and p_2, respectively. (Anderson-Darling test: P-values $p = 0.69$ for p_1 and $p = 0.27$ for p_2. Maximum-likelihood estimators for the mean value m and the standard deviation of the logarithms: $m = -0.71$, $\sigma = 0.48$ for p_1, and $m = -1.3$, $\sigma = 0.63$ for p_2.).

For the quadratic model and the 26 stations with sufficient data for this model, we also considered the roots as potential indicators. However, there were several drawbacks. First, for 25 of 26 stations (exception: 2488), the roots had no 95 %-significant positive or negative (Pearson) correlation with the leading coefficients. Second, at 16 of 26 stations, at least one of the seven roots occurred before 2012 or after 2020. Thus, in general the roots were extrapolations and not directly observable from the data. Third, there was a high variability. For the sample of the 26 stations, the logarithms of average lengths of the 95 %-confidence intervals of the seven roots at each station were approximately normally distributed (mean 4.0, standard deviation 1.92, Anderson-Darling test: P-value $p = 0.39$). According to this distribution, with 81 % probability, for similar data to ours the average lengths of the confidence intervals of the roots would be 10 years or more (while there were at most 9 years of observation).

3.4. Ordinal comparison and pattern of pollution evolution

For each station, there was only one dominant (inside or outside) pollutant. However, owing to the variability in the parameters, another pollutant could dominate by chance (Table 1: simulations). Here, we considered a weakening of this notion, inside pollutants with relatively large leading coefficients. By this we mean coefficients that were above the median of all leading coefficients for the different (inside) pollutants at a given station. The outside pollutant was ignored.

Considering both models (Table 2), pH had 12 times a relatively large coefficient, followed by N (9), DO and TC (8), Con and FC (7), and BOD (6). The three pollutants with relatively large leading coefficients for the best-fit parameters had again relatively large coefficients in 15–73 % of the simulations (Table 2); all other combinations occurred less frequently. Further, comparing at each station the simulated leading coefficients with the best-fit leading coefficients then with one exception (station 2489 with 61 % for the quadratic indicator), in at least 82 % of the 10,000 simulations the Blomqvist correlation was high, $\beta \geq 1/3$. (For this β value, in the simulations at most one of the pollutants could drop out of the group with relatively large indicators.) Hence, we expected to draw more stable conclusions from the relatively large model parameters. We used this notion in our search for common patterns in the dynamics of the evolution of the pollution shares across different stations.

A related notion (comparison with 0) is studied in business literature, where the signs of the interaction coefficients (of the inside goods) of the

Table 1
Dominant pollutants in the best-fit model and in simulations.

Location	Code	Linear LV-models			Quadratic LV models		
		Best fit	Sim: 1st	Sim: 2nd	Best fit	Sim: 1st	Sim: 2nd
Garhmukteshwar, UP	1062	NA			TC	66 % TC	25 % FC
Anoopshahar, UP	2489	Temp	92 % Temp	5 % DO	N	41 % N	24 % Temp
Narora, UP	1145	Temp	72 % Temp	15 % TC	NA		
Kanpur, UP	1067	NA			FC	49 % FC	43 % TC
Kanpur, UP	1068	Temp	38 % Con	32 % Temp	NA		
Kalakankar, UP	2498	Con	47 % Con	16 % Temp	NA		
Allahabad, UP	1046	pH	64 % pH	23 % N	pH	35 % pH	24 % DO
Allahabad, UP	2487	NA			pH	63 % pH	23 % Temp
Allahabad, UP	1049	NA			N	45 % N	36 % DO
Mirzapur, UP	2486	TC	64 % TC	35 % FC	NA		
Mirzapur, UP	2485	NA			N	68 % N	30 % Temp
Tribeni, WB	2506	pH	55 % pH	23 % N	NA		
Kolkata, WB	1054	NA			TC	67 % TC	23 % pH
Kolkata, WB	1053	pH	53 % pH	35 % N	NA		
Kolkata, WB	1470	pH	46 % pH	44 % N	NA		
Diamond Harbor, WB	1469	pH	99 % pH		pH	56 % pH	29 % TC

Note. The column "Best fit" lists the pollutants, which the best-fit model forecasted to finally dominate. The next two columns were based on 10,000 simulations. They inform about the most and second most frequent outcome with respect to domination. The simulated model parameters were random variates from multinormal asymptotic distributions of the best-fit parameters. The distributions of different pollutants were assumed to be independent.

Table 2
Inside pollutants with relatively large coefficients in the best-fit model and in simulations.

Location	Code	Linear LV-model			Quadratic LV-model		
		Best fit	Sim	$\beta \geq 1/3$	Best fit	Sim	$\beta \geq 1/3$
Garhmukteshwar, UP	1062	NA			FC, N, TC	73 %	94 %
Anoopshahar, UP	2489	BOD, DO, pH	33 %	86 %	BOD, N, pH	15 %	61 %
Narora, UP	1145	DO, FC, TC	27 %	82 %	NA		
Kanpur, UP	1067	NA			Con, FC, TC	36 %	85 %
Kanpur, UP	1068	Con, DO, TC	33 %	92 %	NA		
Kalakankar, UP	2498	BOD, Con, N	30 %	84 %	NA		
Allahabad, UP	1046	DO, N, pH	52 %	98 %	BOD, DO, pH	34 %	97 %
Allahabad, UP	2487	NA			DO, N, pH	34 %	94 %
Allahabad, UP	1049	NA			DO, N, pH	45 %	97 %
Mirzapur, UP	2486	BOD, FC, TC	41 %	93 %	NA		
Mirzapur, UP	2485	NA			Con, DO, N	39 %	90 %
Tribeni, WB	2506	FC, pH, TC	60 %	99 %	NA		
Kolkata, WB	1054	NA			FC, pH, TC	68 %	98 %
Kolkata, WB	1053	Con, N, pH	55 %	95 %	NA		
Kolkata, WB	1470	Con, N, pH	73 %	99 %	NA		
Diamond Harbor, WB	1469	BOD, Con, pH	57 %	99.9 %	FC, pH, TC	46 %	98 %

Note. The column "Best fit" lists the inside pollutants with relatively large leading coefficients (alphabetic order); "Sim" counts, how often in 10,000 simulations exactly these three pollutants had relatively large leading coefficients; "$\beta \geq 1/3$" counts, how often the Blomqvist beta was 1/3 or higher. The simulated model parameters were random variates from multinormal asymptotic distributions of the best-fit parameters. The distributions of different pollutants were assumed to be independent.

LV-system characterized its dynamics through an ecological interpretation (Fu et al., 2017): competition (+ +), mutualism (−−), or predator–prey dynamics (+−). For the ten stations where the linear LV-model had acceptable fits, there were three stations with all signs "−" (pairwise mutualism), for two stations, one sign was "−" and the others were "+" (one sole prey for six pairwise competing predators), and for the other stations, all three types of dynamics occurred. For the quadratic LV-model, the signs could change. Here, the leading coefficient would inform about the sign in the far future. As to limitations, in total only 18 (26 %) of 70 leading coefficients of the linear model (10 stations) and 15 (24 %) of 63 leading coefficients of the quadratic model (9 stations) had significant signs (positive lower or negative upper bound of the 95 % confidence interval). Further, the relation to dominance was weak. For the linear model there were only two dominating inside pollutants with significantly positive leading coefficients. (At station 2506, BOD and TC had significantly positive leading coefficients, but pH with insignificant leading coefficient dominated.) For the quadratic model, none of the dominating pollutants had a significantly positive leading coefficient.

We hypothesized that despite the above noted uncertainty about dominance, two stations with a similar long-term evolution of the pollution shares might have similar leading coefficients of the LV-models, at least with respect to their ordinal information relative to the median. (For example, this information would not change, if the pollutant with the second largest leading coefficient would dominate by chance.) Further, we hypothesized that neighboring stations would display a similar long-term evolution of the pollution shares. To test this hypothesis for each of the indicators (a), (b), and (c), at each station we collected the indicator values for the seven (inside) pollutants into a vector in seven dimensions and tested for significantly positive correlations between these vectors. Thereby, we used only the ordinal information (to take care of possible uncertainties) and distinguished only between relatively large and small indicator values (above/below the median). Blomqvist beta by design used exactly this information. Owing to the small number of only seven inside pollutants per stations, we could expect only weak (90 %) significance. (Using the Blomqvist beta test, the observed P-values were 0, 0.1, and 1.) To obtain nevertheless statistically relevant information, we considered all 28 stations with sufficient data for modeling, regardless of the goodness of fit. To visualize the 351 comparisons for each two of the 27 stations with sufficient data for the linear LV-model and the 325 comparisons of 26 stations for the quadratic LV-model, we plotted correlograms: We represented the stations by vertices (points) and linked two of them with an edge (line) if the indicator-vectors had a positive and 90 % significant beta (in dimension 7 this was equivalent to $\beta \geq 1/6$). The vectors were comprised of (a) the linear, (b) the quadratic, and (c) the root indicators associated to the seven inside pollutants. The correlograms are provided as Supporting Data (file SD3).

For all three types of indicators, the correlograms displayed clusters (highlighted in file SD3) that spread across the study area: Regardless of their distance (along river Ganga) every-two stations of the cluster were linked (graph theoretical cliques). The clusters for the linear, quadratic, and root indicators, were comprised of 16 (8 UP, 8 WB, 60 % of 27 stations), 13 (8 UP, 5 WB, 50 % of 26), and 10 (6 UP, 4 WB, 38 % of 26) stations respectively. To check the significance (if similar outcomes could come from arbitrary indicator values), we compared the outcome with a random graph. (We conducted 10,000 simulations of 26 stations, each with seven random numbers as simulated indicators. For these vectors we defined a graph, linking simulated stations if $\beta \geq 1/6$, meaning significantly positive correlation. This was equivalent to choosing a random graph from a Bernoulli graph distribution with 26 vertices and edge probability of 50 %.) The typical clique of the random Bernoulli graphs (40 % of simulations) had size 8. Cliques of size 10 (as for the root indicator) occurred frequently, too (9 % of the simulations). However, cliques with 13 or more elements were rare (0.1 %). Thus, clique size was significant for the linear and quadratic indicators. Further, each station had 90 % significant positive correlations with 7–18 (Median 15), 8–19 (Median 14), and 9–20 stations (Median 13.5), respectively (vertex degree), each station could be linked to any other station using 2–4 edges (vertex eccentricity), and for all three correlograms 53–54 % of all possible links were realized (graph density). However, these values did not differ significantly from the Bernoulli graphs. (For them, median density was 50 %.) Despite this high level of connectedness, for the linear indicators the removal of three stations (1067, 1068 of Kanpur, UP, and 2485 of Mirzapur, UP) would disconnect the correlogram, leaving another cluster (6 UP, 2 WB). This small number was notable: For the quadratic and root indicators, at least 7 or 9 of 26 stations would be needed to disconnect the graphs (vertex cuts). For the Bernoulli graphs, typically 7 vertices were enough to disconnect the graphs (26 %), while only 0.2 % of the random graphs were disconnected by 3 or fewer vertices.

Contrary to our initial assumption, graphs were more disconnected, when only "neighboring" stations with 90 % significant positive beta were linked, removing all indirect links from the correlogram. (We stretched the meaning of "neighbor": It is the next of the considered stations along river Ganga, regardless of the distance.) For the linear indicator, there were 14 components for 27 vertices, for the quadratic indicator 9 components for 26 vertices, and for the root indicator 8 components for 26 vertices (Table 3). For random graphs, 13 components occurred most frequently (13 %), 9 or fewer components occurred for 5 % of simulations and 8 or fewer components for 2 % of simulations. Further, the largest components were comprised of four, nine, and twelve stations for the linear, quadratic, and root indicators, respectively. For the random graphs, maximal components of size four occurred most often (29 %), while components of size nine or larger occurred for 4 %, only. Thus, with respect to the count and size of components, the findings for the root indicator and (size only) the quadratic indicator differed significantly between our study area and the random graphs.

Initially, we expected physical reasons for graph-disconnections. (For very distant stations, or if water quality changes at the confluence of two rivers, or if megacities pollute the river, then the growth characteristics of the stations might become different.) However, we found only one station (2511 in Kolkata, WB), where the connectedness

Table 3
Connectedness components of the adjacency graphs of successive stations.

Location	Code	Component no			Location	Code	Component no		
		L	Q	R			L	Q	R
Garhmukteshwar, UP	1062	1	1	1	Allahabad, UP	1049	9	2	2
Anoopshahar, UP	2488	2	1	1	Mirzapur, UP	2486	10	3	2
Anoopshahar, UP	2489	3	1	1	Mirzapur, UP	2485	10	4	3
Narora, UP	1145	4	1	1	Varanasi, UP	1071	10	4	3
Kachhla, UP	2490	5	1	2	Berhampore, WB	1080	10	4	4
Kannauj, UP	1063	NA	1	2	Tribeni, WB	2506	11	5	5
Kannauj, UP	1066	6	1	2	Kolkata, WB	1054	12	5	5
Bithur, UP	1146	6	1	2	Kolkata, WB	1472	12	NA	NA
Kanpur, UP	1067	6	1	2	Kolkata, WB	1053	12	NA	NA
Kanpur, UP	1068	6	2	2	Kolkata, WB	2511	13	6	6
Dalmau, UP	1147	7	2	2	Kolkata, WB	1471	14	7	7
Kalakankar, UP	2498	7	2	2	Kolkata, WB	1470	14	7	7
Allahabad, UP	1046	7	2	2	Uluberia, WB	1052	14	8	7
Allahabad, UP	2487	8	2	2	Diamond Harbor, WB	1469	14	9	8

Note. Adjacency graphs were defined from positive and 90% significant Blomqvist beta between the indicator sets L, Q, R. These refer to the linear, quadratic and root indicators, defined from the leading coefficients of the linear and quadratic LV-models and from the roots of the interaction coefficients of the quadratic LV-model. NA indicates that the linear/quadratic LV-model was not fitted due to lack of data (after removal of potential outliers).

components for all indicators consisted of this station, only. At another station (1080 in Berhampore, WB) the connectedness components of all indicators ended. (The confluences with rain-fed Mor and Ajay rivers between Berhampore and Tribeni might explain this.) Thus, using different LV-models or different indicators derived from the same model may lead to different perceptions about the pollution growth pattern. However, for our study area (Fig. 1) Table 3 also displays several stretches along river Ganga, where all three indicators agreed about the similarity of the pollution growth patterns (Kannauj to Kanpur, Dalmau to Allahabad, Mirzapur to Varanasi, and part of Kolkata). This suggested similarities between the temporal evolution of the pollutants that may be inherent to the data.

In a similar way, we explored certain associations between pollutants. Thereby, for each pollutant we defined vectors from the indicator values of that pollutant at the considered stations. Again, we used Blomqvist beta. In view of the length of the compared vectors, we could work with 95 % significance. Fig. 7 illustrates, how the different importance-growth indicators identified different patterns for possible interactions between the pollutants. The outcomes depended also on the samples. For the smaller ones, some associations were not observed sufficiently often to be significant (associations of BOD with FC, N, and pH, DO with pH). Conversely, the larger samples might have blurred

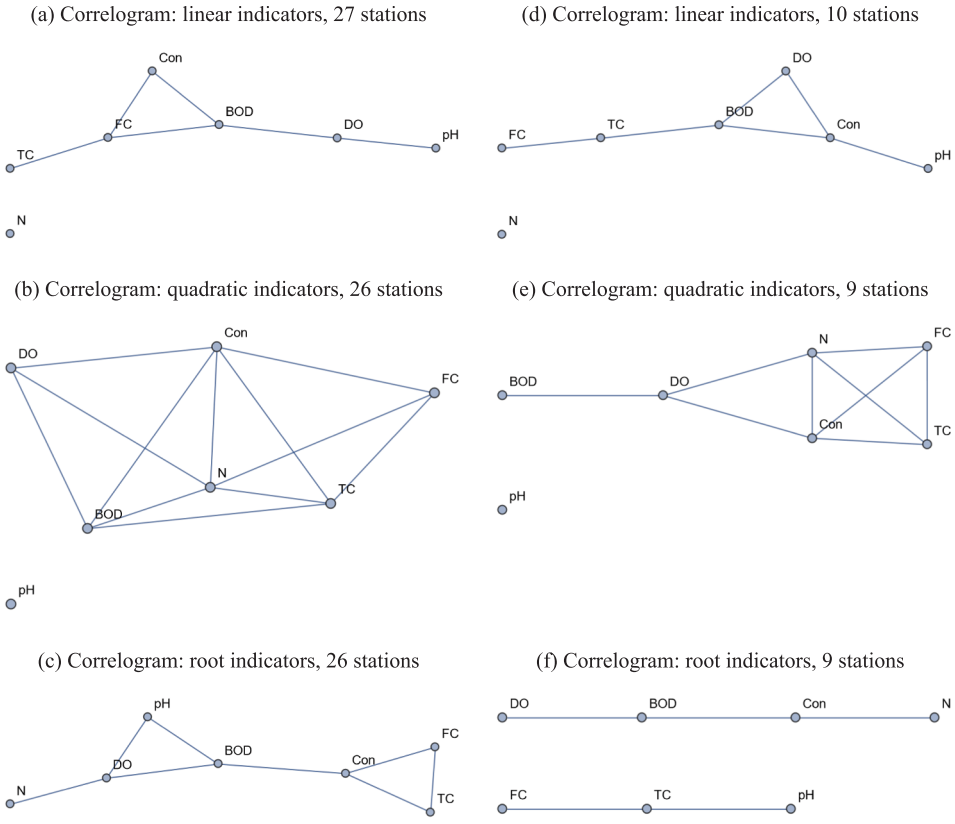


Fig. 7. Pollutants were associated in case of a 95% significant (positive or negative) Blomqvist beta between the vectors of (linear, quadratic and root) indicators of that pollutants at the considered stations. Plots a to c used all available stations and plots d to e used only stations, where the models had acceptable fits. Plot using Mathematica 13.2 and the data from SD2.

some observations that a smaller sample using better fitting models could discern (associations of pH with Con and TC). Further, some associations noted in literature were not replicated for our data (e.g., associations of pH with FC and N). Note that there is a large body of literature on the correlation between pollutants (e.g., Kothari et al., 2021), whereby the focus is on physio-chemical causations. (For instance, regression curves between BOD and chemical oxygen demand COD were derived to estimate BOD more easily: Bhat et al., 2003; Osuide et al., 2011.).

Table 4 lists references that may matter in the present context. Our approach differed insofar from this literature, as we studied the temporal evolution of pollution at given sites, where associations could as well be influenced by social and institutional factors.

4. Discussion and conclusion

4.1. Further discussion of model assumptions

As mentioned in the methods, we have made several assumptions that in hindsight may need more justifications. First, the choice of our data. Based on annual worst-case data, we studied the pollution of river Ganga during 2012–2020, using data from 105 measurement stations. In an initial step, we reduced the count of 105 considered stations to 28, where sufficiently long time-series were available to allow for meaningful tests for the goodness of fit (to refute models with inadequate fit). We thereby discarded the river Ganga data of three states (BR, JH, and UK). This situation was not exceptional: The unavailability of water quality data and the short timespans of observation for the available data are common problems in water resources management (e.g., Cigizoglu and Kisi, 2005; Hadjisolomou et al., 2021; Jeong et al., 2005; Moustaka-Gouni et al., 2019). Thus, we did not find a suitable alternative, namely open data with many monitoring stations (requiring a large river) reporting longer time-series (e.g., monthly data over several decades). For example, we considered other large rivers as possible study sites, such as River Danube or River Rhine in Europe, both with trans-National Monitoring Networks. Thereby, Danube data were not in the public domain and for Rhine, some pollutants were measured at fewer than nine stations.

For the definition of pollution shares we needed an additive water quality index. (This means, WQI was a sum of indices WQ_i, one for each pollutant, i. We then could define the pollution shares as $s_i = WQ_i/WQI$. Thereby, WQ_i needed not be linear, as illustrated for pH.) As our data source focused on a short set of key pollutants, we were restrained to define indices for these pollutants, only. For future research (not only ours) a more complete water quality monitoring along river Ganga might be useful. Several common water quality measurements, such as total suspended sediment (TSS) or heavy metals, were not provided by our source. For an assessment of sustainability, also pesticides, microplastics, or even water-related indicators for land use could be relevant (Yao et al., 2023). Emerging issues with new pollutants or common pollutants that become more important cannot be identified and assessed if monitoring does not consider them. (For a further discussion of water quality assessment: Starkl et al, 2022.).

Table 4
Count of associations observed in the correlograms of Fig. 7.

Association	Count	Association	Count	Association	Count
BOD with DO	6[1]	N with DO	3	BOD with N	1
FC with TC	6[2]	BOD with TC	2	Con with pH	1
BOD with Con	5[1]	DO with pH	2	pH with TC	1[5]
Con with FC	4	FC with N	2[3]	DO with FC	0
Con with DO	3[1]	N with TC	2[3]	DO with TC	0
Con with TC	3	BOD with pH	1	FC with pH	0[5]
N with Con	3	BOD with FC	1[4]	N with pH	0[6]

Notes. Various authors noted similar associations in different contexts. 1: Bhaskar and Dixit, 2015; 2: pollutants originated from the same sources; 3: Aram et al., 2021; 4: Hiraishi et al., 1984; 5: Pearson et al., 1987; 6: Marques et al., 2006.

However, the key pollutants were sufficient to illustrate our method. (Too many additional pollutants could become problematic because of additional potential outliers.) We used a straightforward approach to define our water quality index, relating pollution levels to the pollution thresholds mentioned in the data-source. Apparently, the thresholds of the data-source were intended to inform about water quality for bathing. Did the choice of specific thresholds affect our importance-growth indicators? For instance, using another threshold m_1 for FC defined the index $WQ_1(x_1) = x_1/m_1$ resulting in (dis)utility $u_1(x_1) = \ln(x_1/WQ_0(x_0)) - \ln(m_1)$. Hence, changing m_1 altered the coefficient p_0 of the regression polynomial fitted to the (dis)utility for FC, therefore the root indicator, but not the leading coefficients, therefore not the linear or quadratic indicators. (The same for WQ_2 toWQ_5, and the same for criteria weights, which we did not use.).

The set up of our models involved the definition of an outside pollutant. We used Temp as outside pollutant for all stations and models. To some extent this choice was arbitrary. Selecting a suitable outside pollutant for each station separately could improve the performance of the LV-model (Bauer et al., 2022). However, to define an indicator in a uniform way from the (dis)utilities, we had to choose the same outside pollutant for all stations. Thereby, Temp fulfilled the intuition of an outside good as representing some external remainder (Focacci and Quintavalla, 2020): Its pollution shares were small, and its values were less affected by water management decisions than other pollutants. Conceivable anthropogenic influences on Temp were from the discharge of waste heat of a nuclear power plant at 1145 (Narora, UP), or from dams (e.g., Pashulok Barrage and Bhimgoda Barrage in UK, Farakka Barrage in WB) that may increase Temp (Kędra and Wiejaczka, 2018).

Although the timeseries provided by the data were quite short, we removed additional data points with poor fit. Our rationale was the identification and removal of outliers. For a single timeseries, the judgment of the researcher who collected the data could be a sensible method to identify outliers. For 225 timeseries we preferred an automated approach to ensure transparency and replicability. Thereby, we removed all years, where we detected a potential outlier for one timeseries (at the given station), as we wanted a uniform treatment of all timeseries at a given station. Further, we removed different potential outliers for the linear LV-model and the quadratic LV-model. We were prepared to remove at most one potential outlier per timeseries, whence some outliers might have remained in the selected data, resulting perhaps in poor fits. Indeed, removing all potential outliers from both the quadratic and the linear LV-model from the data would have identified more stations, where one of the models had an acceptable fit.

We fitted only linear and quadratic polynomials to the (dis)utilities. Of course, other function classes could be used for nonlinear regression, too. For example, in energy studies, where the data showed a periodic structure (e.g., business cycles), Fourier polynomials were used for regression (Dominioni et al., 2019). However, more complex functions allow to optimize more parameters, whence fitting such functions to a small dataset may result in overfitting (Granger and Newbold, 1974). This could easily be the case for five to nine datapoints, as in the present situation. (To avoid evident overfitting for the linear and quadratic polynomials, we used them only with at least four or five datapoints, respectively.) Therefore, we did not use more complex functions. Further, we choose these functions, as their extrapolations to infinite time would clearly indicate (perhaps exaggerate) the trend for the evolution of the relative importance (because the polynomials then approached $\pm\infty$). Hence, we expected that the parameters of these functions could be used to define indicators for the evolution of relative importance growth.

In a final step we selected 16 stations, where one of our two LV-models had acceptable fits. Our thresholds to select these models were not restrictive; this may have contributed to the variability of the indicators. However, as the thresholds had to be fulfilled by eight model

curves simultaneously, and as there was already a high variability in the data, using too stringent thresholds might lead to the rejection of all stations. Further, for statistical evaluations and to explore, if a good fit was really needed, we considered also the 28 stations with sufficient data for our models, but perhaps not so good fits. As for an example, the plots in Fig. 7 compare the conclusions that could be drawn using the stations with acceptably fitting models alone (right hand side) and using all stations (left hand side).

4.2. Conclusions for water management

We proposed three indicators to characterize the dynamics of the pollution shares: linear, quadratic, and root indicator. Their definitions were motivated by the quest for methods to recognize the potential success or failure of water management projects (and subsequently also of programs and policies) in real time. For, the success or failure of new infrastructure and improved management practices may evolve too slow to be clearly discernible from measurements of water pollution. Indeed, a linear regression for *WQI* confirmed our perception that during 2012 to 2020 there was not much change in overall Ganga water quality (Section 3.1). Further, emerging challenges for wastewater management may remain undetected because of still low pollution levels. The new indicators aim at addressing this problem, as they focus on the long-term temporal evolution and not on current pollution levels.

We found no viable associations of the root indicator with the long-term performance of pollutants. For the linear and quadratic indicators there was an obvious relation with the temporal evolution of the pollution shares. (The inside pollutant with largest positive indicator value dominates: its share is 100 % in infinite time. If all indicator values are negative, then the outside pollutant dominates.) As a cautionary note, dominance may have different reasons, whence the observation of dominance can only be a first step, followed by a case-by-case assessment (for which we have no data). For example, at three stations the outside pollutant, Temp, was dominant for the linear indicator (confirmed for 32–92 % of the simulations: Table 1). This could indicate a success of wastewater treatment (inside pollutants becoming less important), but it could also be related in another way to water management (such as discharge of waste heat at station 1145) or to climate variations (e.g., failure of monsoon in years 2014 and 2015).

For another example, the indicators for N were of particular interest: In India continuing urbanization may result in a lower fraction of the population choosing a vegetarian diet (Pandey et al., 2020), which may lead to higher levels of N in the wastewater of urban agglomerations. Further, in this case there may be a growing demand for, e.g., poultry. If (backyard) poultry farms discharge (insufficiently treated) sewage into rivers, then this may affect the concentration of N in rural wastewater (poultry litter contains high levels of N), whereby the rapid growth of poultry husbandry in India (Rath et al., 2015) may translate into increasing pollution shares for N. In turn, this might be reflected by higher indicator values for N. Indeed, we found three stations with a positive and maximal indicator for N, and seven stations with above the median indicator values for N, amongst them all three stations at Allahabad, UP. (In Google Maps we found 20 + poultry farms in the area.) Thus, N may become an issue in the future (upgrading of current wastewater treatment plans by an additional purification step, as is common in Europe for N), although since 2012 at all 28 stations the concentrations of N have remained below the legal threshold. As a caveat, forecasts into the infinite future are rather extreme and therefore not always reliable. For example, in business applications it was observed that innovative products often started with low market shares as predators on established products (Bauer et al., 2022). Thereby, the LV-model could easily overestimate the growth potential, as it was fed only with data about the initial phase of rapid growth. Consequently, dominance (limit 100 % in infinite time) indicated a potential for higher relative importance in the future, but the extrapolation to infinite time was merely speculative.

The large spread of the data resulted in a high variability of the model parameters. Depending on the station and model, in up to 65 % of 10,000 simulations (exploring the impact of data variability) the dominant pollutant of the best-fit model was not dominant for the simulated model (Table 1). To overcome the resulting uncertainty, we proposed an ordinal approach, comparing the position of indicator values relative to their median. To further illustrate this concept, we explored the spatial dimension of the temporal evolutions and compared the indicator values at different stations by means of correlograms that linked stations with 90 % significant positive Blomqvist beta between their indicator sets. For all three indicators, the correlograms were connected. Further, we found several characteristics, where one or more of these graphs differed significantly from a random correlogram of 26 simulated stations, whose indicator values were random numbers. Notable features were the larger size of the cliques (linear and quadratic indicators) and the small size of the vertex-cut set (linear indicator). Further, considering only correlations in the flow direction, we noted the small number of connectedness components (root indicator), and the large size of the maximal components (quadratic and root indicators). This indicated similarities in the overall pattern of the dynamics across the study area, but dissimilarities along the direction of the flow. In the latter case, we observed also stretches, where all three indicators showed significant positive correlations for neighboring stations (Table 3). This indicates a somewhat similar pattern of the growth dynamics at least across certain stretches.

In view of the significant differences to a random graph, we hypothesize that the similarities of growth patterns might be related to similar future pollution risks at the study sites (e.g., pressures towards more pollution, similar practices for pollution abatement), even if their current pollution problems were different. Therefore, with respect to water management we suggest that such patterns may inform the new concept of FlexiBAT (Starkl et al., 2018b). It defines benchmarks for wastewater treatment plants from the comparisons of plants with similar site-specific characteristics. For example, if the new indicators forecast an emerging problem with a pollutant, then according to FlexiBAT the concerned communities might require an upscaling of wastewater treatment infrastructure. For other communities, this may be superfluous. Basically, we argue that different (disconnected) stretches of the river may have different risks with water quality in the future, reflected by the connectedness components of the correlograms, whence FlexiBAT may be defined differently, too.

Concerning the implementation of the indicators by practitioners we note that at the station level most computations could be done using a spreadsheet. This is a main advantage of using LV-models of the type proposed by Marasco et al. (2016). The computation of the water quality index, the pollution shares, and the (dis)utilities is straightforward. Linear and quadratic regression is available, too (LINEST function of Excel). The identification of outliers could be done by visual inspection. The subsequent definition of the LV-model functions for the pollution shares is straightforward, too. Simulations, graph theoretical analyses, and statistical tests were easier in Mathematica.

CRediT authorship contribution statement

Norbert Brunner: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – original draft, Writing – review & editing. **Sukanya Das:** Conceptualization, Methodology, Validation, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Markus Starkl:** Conceptualization, Validation, Investigation, Data curation, Writing – original draft, Writing – review & editing, Project administration.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Brunner reports a relationship with European Commission that includes: funding grants. Das reports a relationship with India Ministry of Science & Technology Department of Biotechnology that includes: funding grants. Starkl reports a relationship with European Commission that includes: funding grants.

Data availability

Data will be made available on request.

Acknowledgements

The authors are grateful for the comments by three reviewers that allowed us to improve the paper substantially. The idea for the paper occurred during the project SARASWATI 2.0, "Identifying Best Available Technologies for Decentralized Wastewater Treatment and Resource Recovery for India". Access to the data was obtained through this project, too. This project was jointly funded within the framework of the EU-India water co-operation by the European Union (Horizon 2020 Research and Innovation Program, Grant Agreement no. 821427) and by the Government of India (Department of Science and Technology/ Department of Biotechnology, sanction order DST/IMRCD/India-EU/ Water Call2/SARASWATI 2.0/2018(G)). Open access funding was provided by the University of Natural Resources and Life Sciences, Vienna (BOKU).

Appendix A. Supplementary data

The raw data are provided as a supplementary Excel File SD1.xlsx: It informs about the station code and the year (columns 1-2), and about the pollution data, namely (depending on the needed input for our water quality index) the maximal and/or minimal annual values of pollution parameters (columns 3-11). The data were retrieved from CPCB (2023). Further (column 12), it informs, which data were removed from our study, and why. For the retained data it tabulates the water quality index (columns 13-21), the pollution shares (columns 22-29), and the (dis) utility of pollutants (columns 30-36). It then explains, separately for the linear and the quadratic LV-model, which data were removed because of potential outliers (columns 37-38). The (intermediate) results of the computations for 28 initially selected stations are provided as a supplementary Excel File SD2.xlsx: It numbers the data in the direction of flow, provides village names, station codes and the name of the pollutant, whose time-series was analyzed (columns 1-4). Further, for each station and model it reported if the model could be accepted or should be rejected (explaining the reason: misspecification or poor fit). A Word file SD3.docx contains the plots of three correlograms linking stations with a similar pattern of importance growth, based on three indicators (Section 3.2) to describe the dynamics of this growth.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecolind.2023.110201.

References

Abbasi, T., Abbasi, S.A., 2012. Water Quality Indices. Elsevier, Amsterdam, Netherlands,. https://doi.org/10.1016/C2010-0-69472-7.

Alley, K.D., 2019. River Goddesses, Personhood and Rights of Nature: Implications for Spiritual Ecology. Religions 10, 502. https://doi.org/10.3390/rel10090502.

Aram, S.A., Saalidong, B.M., Osei Lartey, P., 2021. Comparative assessment of the relationship between coliform bacteria and water geochemistry in surface and ground water systems. PLoS One 16, 0257715. https://doi.org/10.1371/journal.pone.0257715.

Bass, F., 1969. A new product growth for model consumer durables. Manag. Sci. 15, 215–227.

Bauer, R., Schwarzmayr, F., Brunner, N., Kühleitner, M., 2022. Dynamics of the Austrian Food Market: Application of Lotka-Volterra Differential Equations. Open Journal of Modelling and Simulation 10, 152–164.

Beck, M.W., Mazor, R.D., Theroux, S., Schiff, K.C., 2019. The Stream Quality Index: A multi-indicator tool for enhancing environmental management. Environmental and Sustainability Indicators 1, 100004. https://doi.org/10.1016/j.indic.2019.100004.

Bharti, N., Katyal, D., 2011. Water quality indices used for surface water vulnerability assessment. Int. J. Environ. Sci. 2, 154–173.

Bhaskar, M., Dixit, A.K., 2015. Water Quality Appraisal of Hasdeo River at Korba in Chhattisgarh, India. International Journal of Science and Research 4, 1252–1258.

Bhat, M.R., Roopali, S.H., Kulkarni, V.R., 2003. Correlation between BOD, COD and TOC. J. Ind. Pollut. Control 19, 187–191.

Bhutiani, R., Khanna, D.R., Kulkarni, D.B., Ruhela, M., 2016. Assessment of Ganga river ecosystem at Haridwar, Uttarakhand, India with reference to water quality indices. Appl Water Sci 6, 107–113. https://doi.org/10.1007/s13201-014-0206-6.

Blomqvist, N., 1950. On a measure of dependence between two random variables. Ann. Math. Stat. 21, 593–600.

Breitenmoser, L., Cuadrado Quesada, G., Anshuman, N., Bassi, N., Dkhar, N.B., Phukan, M., Kumar, S., Naga Babu, A., Kierstein, A., Campling, P., Hooijmans, C.M., 2022. Perceived drivers and barriers in the governance of wastewater treatment and reuse in India: Insights from a two-round Delphi study. Resour. Conserv. Recycl. 182, 106285 https://doi.org/10.1016/j.resconrec.2022.106285.

Brunner, N., Lele, A., Starkl, M., Grassini, L., 2010. Water sector reform policy of India: Experiences from case studies in Maharashtra. J. Policy Model 32, 544–561.

Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., Liu, F., Zuo, M., Zou, X., Wang, J., Zhang, Y., 2020. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. Water Res. 171, 115454 https://doi.org/10.1016/j.watres.2019.115454.

Cid, C.A., Abiola, F., Starkl, M., 2022. Can international non-sewered sanitation standards help solve the global sanitation crises? Environ. Sci. Tech. 56, 699–706. https://doi.org/10.1021/acs.est.1c03471.

Cigizoglu, H.K., Kisi, Ö., 2005. Flow prediction by three back propagation techniques using k-fold partitioning of neural network training data. Hydrological Research 36, 49–64.

CPCB 2023. *NWMP Data*. Central Pollution Control Board, Government of India, New Delhi, NCT, India, Link: https://cpcb.nic.in/nwmp-data/ last visit 01.03.2023.

Dang, H.S., Huang, Y.F., Wang, C.N., Nguyen, T.M.T., 2016. An Application of the Short-Term Forecasting with Limited Data in the Healthcare Traveling Industry. Sustainability 8, 8101037. https://doi.org/10.3390/su8101037.

Diebold, F.X., 2007. Elements of Forecasting. Thomson South-Western, Mason, OH, USA.

Dominioni, G., Romano, A., Sotis, C., 2019. A Quantitative Study of the Interactions between Oil Price and Renewable Energy Sources Stock Prices. Energies 12 published online: DOI 10.3390/en12091693.

Dutta, V., Dubey, D., Kumar, S., 2020. Cleaning the River Ganga: Impact of lockdown on water quality and future implications on river rejuvenation strategies. Sci. Total Environ. 743, 140756 https://doi.org/10.1016/j.scitotenv.2020.140756.

Eck, D.L., 1999. Banāras. Columbia University Press, NY, USA, City of Light.

Focacci, C.N., Quintavalla, A., 2020. Unpredictable spillovers among water uses? An analysis of agricultural, industrial, and household uses of water in the Balkans. PLoS One 15, 0235079. https://doi.org/10.1371/journal.pone.0235079.

Fu, X., Zhang, P., Zhang, J., 2017. Forecasting and Analyzing Internet Users of China with Lotka-Volterra Model. Asia-Pacific J. Oper. Res. 34 https://doi.org/10.1142/S0217595917400061.

Gai, R., Guo, Z., 2023. A water quality assessment method based on an improved grey relational analysis and particle swarm optimization multi-classification support vector machine. Front. Plant Sci. 25, 1099668. https://doi.org/10.3389/fpls.2023.1099668.

GoI 2009. *Status paper on River Ganga, State of Environment and Water Quality*. National River Conservation Directorate at Ministry of Environment and Forests, Government of India, New Delhi, NCT, India.

Granger, C.W.J., Newbold, P., 1974. Spurious regressions in econometrics. J. Econ. 2, 111–120.

Hadjisolomou, E., Stefanidis, K., Herodotou, H., Michaelides, M., Papatheodorou, G., Papastergiadou, E., 2021. Modelling Freshwater Eutrophication with Limited Limnological Data Using Artificial Neural Networks. Water 13, 1590. https://doi.org/10.3390/w13111590.

Hasan, M.M., Ahmed, M.S., Adnan, R., Shafiquzzaman, M., 2020. Water quality indices to assess the spatio temporal variations of Dhaleshwari river in central Bangladesh. Environmental and Sustainability Indicators 8, 100068. https://doi.org/10.1016/j.indic.2020.100068.

Hiraishi, A., Saheki, K., Horie, S., 1984. Relationships of Total Coliform Fecal Coliform, and Organic Pollution Levels in the Tamagawa River. Nippon Suisan Gakkai Shi 50, 991–997.

Horgan, D.C. 2020. *Modelling market share in the UK grocery retail sector using the nonautonomous Lotka-Volterra equations*. Preprint, available from www.researchgate.net.

Horton, R.K., 1965. An index number system for rating water quality. J. Water Pollut. Control Fed. 37, 300–306.

Huang, Q., Lin, Y., Zhong, Q., Ma, F., Zhang, Y., 2020. The Impact of Microplastic Particles on Population Dynamics of Predator and Prey: Implication of the Lotka-Volterra Model. Sci. Rep. 10, 4500. https://doi.org/10.1038/s41598-020-61414-3.

Jeong, K.S., Kim, D.K., Chon, T.S., Joo, G.J., 2005. Machine Learning Application to the Korean Freshwater Ecosystems. Korean Journal of Ecology 28, 405–415.

Jha, B., Jha, M., 2013. Rating Curve Estimation of Surface Water Quality Data Using LOADEST. J. Environ. Prot. 4, 849–856. https://doi.org/10.4236/jep.2013.48099.

Kędra, M., Wiejaczka, Ł., 2018. Climatic and dam-induced impacts on river water temperature: Assessment and management implications. Sci. Total Environ. 626, 1474–1483.

Kloppers, P.H., Greeff, J.C., 2013. Lotka-Volterra model parameter estimation using experiential data. Appl. Math Comput. 224, 817–825.

Kothari, V., Vij, S., Sharma, S.K., Gupta, N., 2021. Correlation of various water quality parameters and water quality index of districts of Uttarakhand. Environmental and Sustainability Indicators 9, 100093. https://doi.org/10.1016/j.indic.2020.100093.

Kouadri, S., Elbeltagi, A., Islam, A.R.M.T., Kateb, S., 2021. Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region Algerian southeast. Applied Water Sciences 11, 190. https://doi.org/10.1007/s13201-021-01528-9.

Marasco, A., Picucci, A., Romano, A., 2016. Market Share Dynamics Using Lotka-Volterra Models. Technol. Forecast. Soc. Chang. 105, 49–62.

Marques, R., Pereira Zamparoni, C.A.G., de Castro e Silva, E., Barbosa, A.M., Arruda, D., Evangelista, S., de Magalhães, A. 2006. *Correlation analyzes between pH values, nitrate concentration, conductivity electric, precipitation volume and wind direction in Cuiabá city rain*, Mato Grosso state, Brazil. Proceedings of 8th ICSHMO, Foz do Iguaçu, Brazil, April 24-28, 2006, INPE, pp. 131-137.

Mirza, M.M.Q., 2004. The Ganges water diversion: environmental effects and implications. Springer, Dordecht, Netherlands. 1–6. https://doi.org/10.1007/978-1-4020-2792-5.

Modis, T., 1999. Technological forecasting at the stock market. Technol. Forecast. Soc. Chang. 62, 173–202.

Moustaka-Gouni, M., Sommer, U., Economou-Amilli, A., Arhonditsis, G.B., Katsiapi, M., Papastergiadou, E., Kormas, K.A., Vardaka, E., Karayanni, H., Papadimitriou, T., 2019. Implementation of the Water Framework Directive: Lessons Learned and Future Perspectives for an Ecologically Meaningful Classification Based on Phytoplankton of the Status of Greek Lakes, Mediterranean Region. Environ. Manag. 64, 675–688.

Newey, W.K., West, K.D., 1987. A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. Econometrica 55, 703–708.

Osuide, M.O., Ezeaku, E., Nwaiwu, P.C., Akeno, S.E., Okojie, V.U., Edokpa, W.I., 2011. Studies on the correlation of some aggregate parameters in the drains of a service facility. Global J. Pure Appl. Sci. 17, 311–318.

Pandey, B., Reba, M., Joshi, P.K., Seto, K.C., 2020. Urbanization and food consumption in India. Sci. Rep. 10, 17241. https://doi.org/10.1038/s41598-020-73313-8.

Parween, S., Siddique, N.S., Diganta, M.T.M., Olbert, A.I., Uddin, M.G., 2022. Assessment of urban river water quality using modified NSF water quality index model at Siliguri city, West Bengal India. Environmental and Sustainability Indicators 16, 100202. https://doi.org/10.1016/j.indic.2022.100202.

Pearson, H.W., Mara, D.D., Mills, S.W., Smallman, D.J., 1987. Physico-chemical parameters influencing fecal bacterial survival in waste stabilization ponds. Water Sci. Technol. 19, 145–152.

Rath, P.K., Mandal, K.D., Panda, P., 2015. Backyard poultry farming in India: A call for skill upliftment. Res J Recent Sci 4, 1–5.

Shamshirband, S., Nodoushan, E.J., Adolf, J.E., Manaf, A.A., Mosavi, A., Chau, K., 2019. Ensemble models with uncertainty analysis for multi-day ahead forecasting of chlorophyll a concentration in coastal waters. Engineering Applications of Computational Fluid Mechanics 13, 91–101.

Starkl, M., Brunner, N., Stenström, T.A., 2013. Why Do Water and Sanitation Systems for the Poor Still Fail? Policy Analysis in Economically Advanced Developing Countries. Environ. Sci. Tech. 47, 6102–6110. https://doi.org/10.1021/es3048416.

Starkl, M., Brunner, N., Hauser, A.W.H., Feil, M., Kasan, H., 2018a. Addressing Sustainability of Sanitation Systems: Can it be Standardized? International Journal of Standardization Research 16, 39–51. https://doi.org/10.4018/IJSR.2018010103.

Starkl, M., Anthony, J., Aymerich, E., Brunner, N., Chubilleau, C., Das, S., Ghangrekar, M.M., Kazmi, A.A., Philip, L., Singh, A., 2018b. Interpreting best available technologies more flexibly: A policy perspective for municipal wastewater management in India and other developing countries. Environ. Impact Assess. Rev. 71, 132–141. https://doi.org/10.1016/j.eiar.2018.03.002.

Starkl, M., Brunner, N., Das, S., Singh, A., 2022. Sustainability Assessment for Wastewater Treatment Systems in Developing Countries. Water 14, 241. https://doi.org/10.3390/w14020241.

Sterman, J.D., 2001. System dynamics modeling: Tools for learning in a complex world. Calif. Manage. Rev. 43, 8–25. https://doi.org/10.2307/41166098.

Tang, M., Xu, W., Zhang, C., Shao, D., Zhou, H., Li, Y., 2022. Risk assessment of sectional water quality based on deterioration rate of water quality indicators: A case study of the main canal of the Middle Route of South-to-North Water Diversion Project. Ecol. Ind. 135, 108592 https://doi.org/10.1016/j.ecolind.2022.108592.

Wohl, E.E., 2012. A world of rivers: environmental change on ten of the world's great rivers. University of Chicago Press, Chicago, IL, USA.

Wolfram Research 2023. *Mathematica, Version 13.2*. Wolfram Research Inc., Champaign, IL. Link: www.wolfram.com/mathematica, last visit 01.03.2023.

Yao, S., Chen, C., He, M., Cui, Z., Mo, K., Pang, R., Chen, Q., 2023. Land use as an important indicator for water quality prediction in a region under rapid urbanization. Ecol. Ind. 146, 109768 https://doi.org/10.1016/j.ecolind.2022.109768.

Ziegler, A.M., Brunner, N., Kühleitner, M., 2020. The Markets of Green Cars of Three Countries: Analysis Using Lotka-Volterra and Bertalanffy-Pütter Models. Journal of Open Innovation: Technology, Market and Complexity 6, 67. https://doi.org/10.3390/joitmc6030067.