


RESEARCH

Open Access



Chemical surveillance in freshwaters: small sample sizes underestimate true pollutant loads and fail to detect environmental quality standard exceedances

Denise Babitsch^{1,2*}  and Andrea Sundermann^{1,2}

Abstract

Background: Chemical surveillance in surface waters is crucial to identify potential threats to the health of freshwater ecosystems. Usually, the concentrations of pollutants are highly variable over the course of the year and often result in non-normally distributed data sets. Therefore, the European Water Framework Directive recommends measuring, e.g. priority substances at least 12 times a year to achieve an acceptable accuracy level for the estimation of the true mean annual loads. However, in Europe priority substances are often measured much less frequently. In this context, the aim of the present study was to analyze how sample size, temporal variability and skewness of the data sets influence the accuracy of the mean annual load estimation and the assessment of annual average environmental quality standards. For this purpose, sample size simulations using weekly composite samples of benzo(a)pyrene, 4-*tert*-octylphenol, fluoranthene and di(2-ethylhexyl) phthalate, selected as representatives for priority substances, were carried out.

Results: The sample size simulations showed two general patterns: the accuracy of the mean annual load estimation increased with increasing sample size and skewness and temporal variability were more apparent in smaller sample sizes. In right-skewed data sets, small sample sizes led, on average, to a systematic underestimation of the true mean annual load whilst in a few cases these led to an overestimation. Although the study was carried out on priority substances, results can be transferable to other pollutants. Furthermore, in small sample sizes a considerable proportion of the simulated means failed to detect annual average environmental quality standard exceedances.

Conclusions: The results of the present study indicate that the usage of small sample sizes is likely to result in an underestimation of the true mean annual pollutant loads in chemical surveillance and scientific research, thus potentially jeopardizing the validity of results. Therefore, it is recommended to avoid the usage of small sample sizes for the determination of mean annual pollutant loads. Furthermore, priority substances should be sampled according to the European Water Framework Directive guidelines at least 12 times/year to improve the assessment of the threat posed by pollutants to freshwater ecosystems in Europe.

Keywords: Micropollutant, Priority substance, Sample size simulation, Monitoring, Weekly composite sample, Accuracy assessment, Uncertainty, European Water Framework Directive, Temporal variability

Background

Freshwater organisms are confronted with numerous stressors in European surface waters [1, 2], of which water pollution is considered to be one of the most significant and widespread [3, 4]. Chemicals, even if they

*Correspondence: Denise.Babitsch@senckenberg.de

¹ Department of River Ecology and Conservation, Senckenberg Research Institute and Natural History Museum Frankfurt, Clamecystr. 12, 63571 Gelnhausen, Germany

Full list of author information is available at the end of the article

typically occur at very low concentrations ranging from ng/L to µg/L (so-called micropollutants), contribute to the loss of freshwater biodiversity [4–6] and are, amongst others, the reason for failing the good ecological status of the European Water Framework Directive (WFD) [1, 3].

Therefore, the surveillance of micropollutants in surface waters is essential. Due to the WFD, extensive chemical monitoring programs which assess the chemical status of surface waters are already available in Europe [7]. The chemical status is defined Europe-wide on the basis of 45 priority substances in terms of their compliance with environmental quality standards (EQS) [8]. Priority substances are (micro)pollutants that are classified as substances that pose a significant risk to the aquatic environment. The EQS for inland surface waters are substance-specific threshold concentrations that should not be exceeded in surface waters in order to protect freshwater organisms and the environment. For the annual assessment of EQS compliance, the mean annual and maximum surface water concentrations of priority substances are used. According to the WFD, priority substances should be sampled at least 12 times a year to ensure a high accuracy of the estimation of the true mean annual load and, thus, reliable surveillance [7]. The reality, however, is that priority substances are often sampled much less frequently [9], probably due to cost efficiency, capacity efficiency and time efficiency reasons. In addition, in many scientific monitoring studies which monitor the occurrence of priority substances on a large scale, the sample frequencies are often significantly lower than 12 samples per year [see the data sets used in 10, 11].

Due to this large-scale use of low sampling frequencies for the calculation of the mean annual pollutant loads, the question must be asked whether small sample sizes are accurate enough to correctly represent the concentrations of priority substances occurring in surface waters in the course of a year. When addressing this question, it is important to bear in mind that the concentration curves of micropollutants can be very variable during the year. This temporal variability is caused by various factors, such as different input paths [12–14], diurnal variability [15, 16], seasonal variability [17, 18] and variable base flow conditions of the surface water [19] which can lead to large concentration ranges. Another point that should be considered is that annual concentration curves of (micro)pollutants usually result in right-skewed data sets [11, 20, 21], which means that they mainly consist of low to medium concentrations and only a few high concentrations. These high, sporadically occurring, concentrations can be caused, for example, by surface runoff [22, 23], stormwater and combined sewer overflows [24–26] due to heavy rainfall events, or by accidental events in surface waters, triggering significant ecotoxicological

effects in the aquatic environment [27, 28]. Due to these effects, the detection of high micropollutant concentrations is of particular importance for the surveillance of surface waters. If micropollutants are not measured sufficiently often during the year, then there is a high probability that these occasional high concentrations will not be detected. In addition to the above question, it is important to know the impact of temporal variability and right-skewness on the estimation accuracy and whether and how a small sample size will affect the accuracy of the EQS assessment.

The best information for answering these questions is provided by high-frequency measurements (i.e. sub-hourly, hourly). However, for micropollutants, such as priority substances, data sets measured at such high resolutions are scarcely available. In Germany, the highest sampling frequency of priority substances is measured by a limited number of permanent monitoring stations as weekly composite samples. Weekly composite samples consist of sub-hourly water samples which are combined into one sample after seven consecutive days. A weekly composite sample reflects the average weekly pollutant load, and 52 weekly composite samples reflect the average course of the annual concentration curve. For this reason, the weekly composite samples provide a good basis to investigate the influence of sample size on the accuracy of the mean annual load estimation and on the EQS assessment. Since the data sets are so highly resolved, it is also possible to estimate the skewness and temporal variability of the micropollutant loads and to include these data set characteristics in the investigations. For this purpose, sample size simulations were carried out with weekly composite samples of priority substances on the basis of a simulation strategy according to Thompson [29]. Benzo(a)pyrene, 4-*tert*-octylphenol, fluoranthene and di(2-ethylhexyl) phthalate (DEHP) were selected as representatives for priority substances.

Thus, the primary objective of the present study was to analyze (a) how sample size and data set characteristics, such as skewness and temporal variability, influence the accuracy of the mean annual load estimation, (b) whether small sample sizes lead to a systematic error (under- or overestimation) of the true mean annual load and (c) whether and how sample size affects the accuracy of the EQS assessment.

Methods

Data and data set characteristics

Due to their high quality and high sampling frequency, monitoring data from the federal state of Saxony (Germany) were used in this study. The Saxon State Office for Environment, Agriculture and Geology provided measured concentrations of more than 500 pollutants.

Substances and data sets were selected for the analyses according to the following criteria:

1. Substances had to be classified as priority substances of the European Directive 2013/39/EU [8],
2. Data sets had to be measured as weekly composite samples with a sampling frequency of 52 measured values per year corresponding to 52 calendar weeks,
3. All 52 data set values had to be above the limit of quantification and
4. The mean annual load of at least one data set per substance had to exceed the respective EQS for inland surface waters [8].

Of the more than 500 substances, only the four substances benzo(a)pyrene, 4-*tert*-octylphenol, fluoranthene and DEHP met the required criteria.

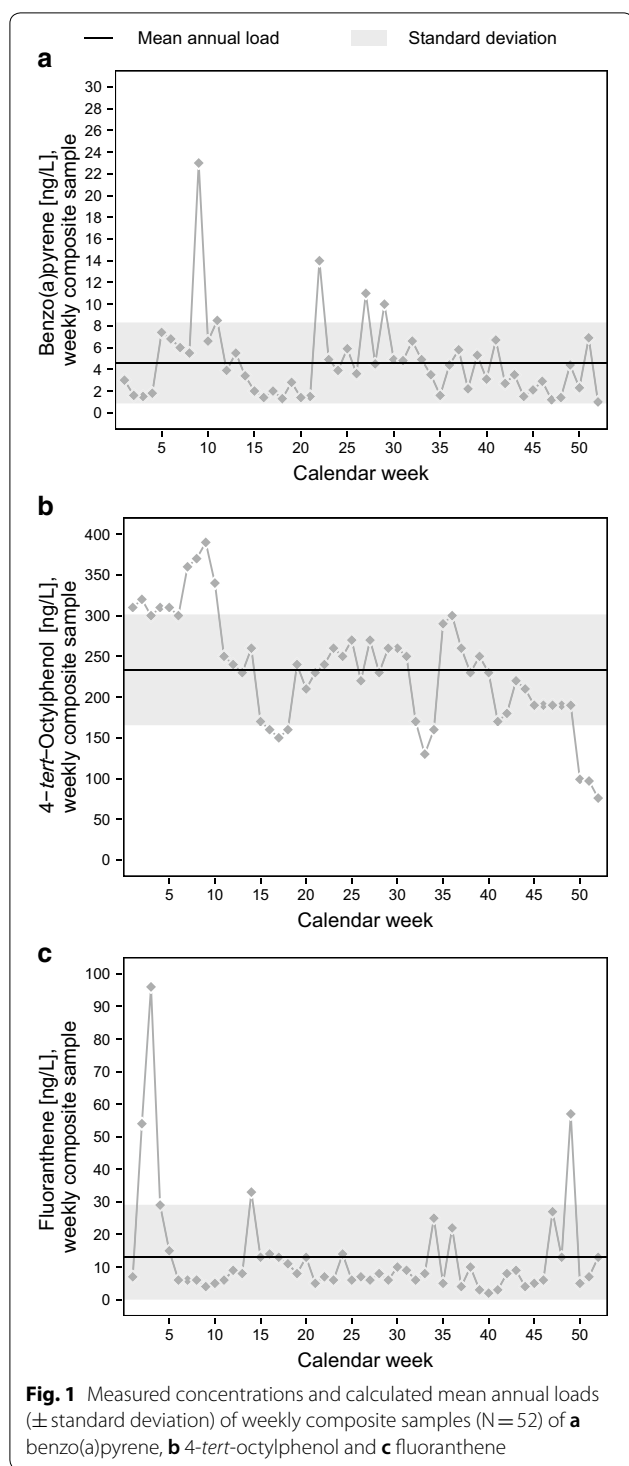
Benzo(a)pyrene and fluoranthene belong to the group of polycyclic aromatic hydrocarbons (PAHs). PAHs occur due to incomplete combustion of organic matter, such as wood, coal or oil. In addition to natural sources of PAHs in the environment, like forest fires [30, 31] and volcanic eruptions [32], there are numerous anthropogenic sources, such as coke oven emissions [33, 34], combustion of fossil fuels [35] and vehicle exhausts [36]. Furthermore, benzo(a)pyrene and fluoranthene are detected in many different products. For example, both substances are plasticizer components in polymers [37], fluoranthene is an intermediate for dyes and pharmaceuticals [38] and benzo(a)pyrene occurs as an impurity product in cooling lubricants [39]. 4-*tert*-Octylphenol is a constituent of phenolic resins and a starting substance for the production of polymers and a group of surfactants, the octylphenol ethoxylates [40]. 4-*tert*-Octylphenol itself, the resins and ethoxylates are, e.g. used in tires, printing inks and veterinary medicine formulations. In addition to the parent substance, the ethoxylates also contribute to the emission of 4-*tert*-octylphenol into the environment, as these are transformed back into 4-*tert*-octylphenol

by degradation processes. As gaseous or particulate bound substances, benzo(a)pyrene, fluoranthene and 4-*tert*-octylphenol are expected to be emitted into surface waters mainly via sewer systems and sealed urban areas, industrial direct dischargers, erosion, municipal wastewater treatment plants and atmospheric deposition [41]. DEHP is a plasticizer, mainly used in the production of polymer products, such as flexible polyvinyl chloride [42]. This substance can be found in building materials, coated fabrics, medical devices and in a wide range of other products. As a predominantly particulate bound substance, DEHP is mainly expected to enter surface waters via sewer systems and sealed urban surfaces, industrial direct dischargers, atmospheric deposition and erosion [41]. Due to its reproductive toxicity and its endocrine disruptive properties, DEHP was identified as substances of very high concern and therefore included in the Candidate List of the European Chemicals Agency in 2008 [42, 43]. Since the sunset date in January 2015, the active substance may only be placed on the market and used with approval [44]. Additional uses and import are heavily restricted.

The final selection of the data sets for the analyses was made from 60 data sets that met the above criteria. In the end, a total of seven data sets from three permanent monitoring stations were selected (Table 1; Figs. 1, 2). For benzo(a)pyrene and 4-*tert*-octylphenol, the most recent available data set was chosen. To investigate the influence of high, sporadically occurring, pollutant concentrations on the sample size simulations, the data set with the most outliers was selected for fluoranthene. Outliers were defined according to Tukey [45]. The selected fluoranthene data set contained a total of seven outliers. To compare the fluoranthene data set with a data set that contained no outliers, a second, manipulated data set was created on the basis of the original fluoranthene data set. In this manipulated data set, all seven outliers were set to the mean annual load of the original data set. Since the inclusion in the Candidate List, the annual pollutant

Table 1 Overview of selected data sets and sampling locations

Substance	Sampling year	Permanent monitoring station	River	Coordinates
Benzo(a)pyrene	2016	Zehren	Elbe	N51° 12' 34", E13° 24' 15"
4- <i>tert</i> -Octylphenol	2012	Schmilka	Elbe	N50° 53' 29", E14° 13' 49"
Fluoranthene	2015	Schmilka	Elbe	N50° 53' 29", E14° 13' 49"
DEHP	2007	Domnitzsch	Elbe	N51° 38' 55", E12° 53' 42"
DEHP	2008	Domnitzsch	Elbe	N51° 38' 55", E12° 53' 42"
DEHP	2010	Domnitzsch	Elbe	N51° 38' 55", E12° 53' 42"
DEHP	2016	Domnitzsch	Elbe	N51° 38' 55", E12° 53' 42"



loads of DEHP have constantly decreased which indicates the success of the regulatory measures. Due to the decreasing pattern of concentrations and annual variability of DEHP over time, the substance is suitable for assessing the influence of temporal variability on sample

size simulations. For the analyses, the years 2007 and 2008 (1 year before and the year of DEHP's inclusion in the Candidate List), 2016 (the year after the sunset date) and 2010 (1 year in between) were selected.

The data set characteristics of all four substances are summarized in Table 2. The distribution of all data sets was checked for skewness and normality (Shapiro–Wilk test, $\alpha=0.05$; Table 2). Benzo(a)pyrene, both fluoranthene data sets, as well as the 2007, 2008 and 2010 DEHP data sets, were right-skewed and not normally distributed. The skewness of these data sets ranged from 0.72 to 3.42. The 4-tert-octylphenol and the 2016 DEHP data set were normally distributed.

Weekly composite samples

Weekly composite samples were measured at the permanent monitoring stations as follows: a water volume of 50 mL was automatically collected every 45 min and combined into a daily composite sample. 300 mL of each of seven consecutive daily composite samples were again combined to yield a weekly composite sample. This was then analyzed in the state office laboratory according to the applicable DIN standards.

Sample size simulation

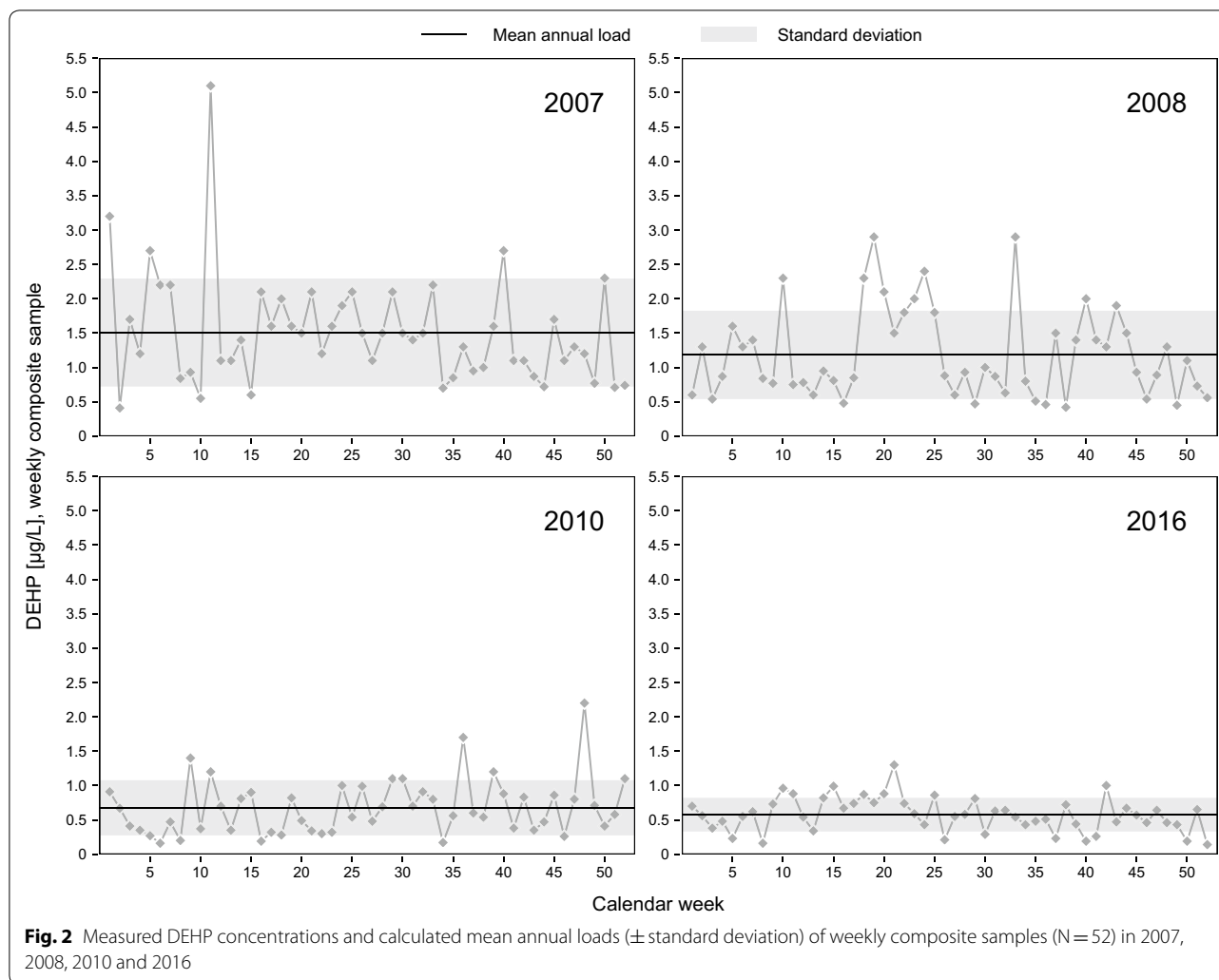
Since the state offices evaluate the chemical water quality based on mean annual loads, annual mean values were selected as analysis parameters in the study. The annual mean value of the 52 weekly composite samples was used as the reference mean value and is, from this point forward, referred to as the “true mean”. For the simulation, simple random sub-samples of $n=1$ to $n=51$ were selected out of the 52 weekly composite samples and the sub-sample means were calculated. The procedure was performed 100,000 times to assess the distribution of the simulated sub-sample means. This distribution was used as an estimator of the true mean.

EQS assessment

The mean annual loads of the benzo(a)pyrene, 4-tert-octylphenol, fluoranthene and 2007 DEHP data sets exceeded the respective annual average EQS (AA-EQS) for inland surface waters [8] (Fig. 3). To investigate the influence of sample size on the EQS assessment, the proportion of simulated sub-sample means below or equal to the AA-EQS was calculated and defined as “false negatives for AA-EQS exceedance”.

Statistical analysis software

Statistical analyses were performed using R, version 3.5.3 [46]. Sample size simulations were run with a script by Thompson [29] which was modified for this study.



Results

Accuracy of the true mean estimation

Figure 4 shows how the 100,000 sub-sample means of benzo(a)pyrene randomly sampled per sub-sample size were distributed around the true mean value. It was found that the values of the sub-sample means approximated the true mean with increasing sub-sample size, resulting in an increase in the accuracy of the true mean estimation. The observed pattern was apparent in all simulations (Figs. 4, 5, 6, 7). This indicates that the more often water samples are taken per year, the more likely it is that the true annual mean value will be correctly estimated.

Sample size and skewness

The sample size simulation of benzo(a)pyrene (Fig. 4) showed that the true mean was underestimated on average (median) for smaller sub-sample sizes. This pattern was also apparent in the fluoranthene sample size

simulation (Fig. 5a) as well as in the 2007, 2008 and 2010 DEHP sample size simulations (Fig. 7). It was found that the underestimation was caused by the right-skewness of the weekly composite sample data sets (see Table 2). Since right-skewed distributed data sets mainly contain concentrations in lower ranges and only to a small extent high concentrations, for smaller sub-sample sizes there is a high statistical probability that only the lower concentrations will be included in a sub-sample. This high probability will lead to the observed average underestimation of the true mean in the sample size simulations. With increasing sub-sample size, the probability of including the few high concentrations in a sub-sample increases. As a result, the majority of the sub-sample mean values increase and the median (see boxplots in Figs. 4, 5, 6 and 7—2007, 2008, 2010) approximates the true mean as the sub-sample size increases.

As already mentioned, in right-skewed distributed data sets, the statistical probability of taking a sub-sample that

Table 2 Data set characteristics of the selected data sets

Data set	Mean annual load	SD	Range	Skewness	Deviation from normality (p value)
Benzo(a)pyrene	4.58 ^a	3.73 ^a	22.0 ^a	2.70	4.55×10^{-8}
4-tert-Octylphenol	233 ^a	68.0 ^a	314 ^a	-0.02	0.89
Fluoranthene	13.1 ^a	16.0 ^a	94.0 ^a	3.41	3.33×10^{-11}
Fluoranthene, manipulated data set	8.71 ^a	4.02 ^a	20.0 ^a	0.72	1.17×10^{-3}
DEHP 2007	1.51 ^b	0.79 ^b	4.69 ^b	1.97	1.02×10^{-5}
DEHP 2008	1.18 ^b	0.64 ^b	2.48 ^b	0.96	2.71×10^{-4}
DEHP 2010	0.68 ^b	0.40 ^b	2.04 ^b	1.35	2.9×10^{-4}
DEHP 2016	0.58 ^b	0.25 ^b	1.16 ^b	0.32	0.43

All calculated values are based on the concentrations of 52 weekly composite samples. Variability is expressed as SD (standard deviation) and range. Normality was checked by the Shapiro–Wilk test ($\alpha = 0.05$)

^a ng/L

^b µg/L

does not contain any of the few high concentrations is greater for smaller sub-sample sizes than for larger ones. Nevertheless, a high concentration in the sub-sample has a much stronger effect on the sub-sample mean value in smaller sub-sample sizes which can lead to a strong overestimation of the true mean. This can be seen, for example, in the maximum sub-sample means of the benzo(a)pyrene sample size simulation (Fig. 4).

To investigate the influence of the few high pollutant concentrations in right-skewed distributed data sets on the sample size simulations, a second fluoranthene data set, the manipulated data set, was created based on the original fluoranthene data set. In this manipulated data set, the seven outliers of the original fluoranthene data set were eliminated by setting them to the original data set's true mean. This created a very slightly right-skewed and approximately normally distributed data set (see Table 2). The sample size simulation of the manipulated data set (Fig. 5b) demonstrated very clearly how the skewness affects the result of sample size simulations, especially in the case of smaller sub-sample sizes. The two patterns, of the average underestimation and of the overestimation of the true mean, were significantly attenuated compared to the sample size simulation of the original fluoranthene data set (Fig. 5a).

The data sets of 4-tert-octylphenol and 2016 DEHP were normally distributed (Table 2). Compared to a right-skewed data set, a normally distributed data set has the highest data density around the mean value. This distribution was clearly reflected in the

corresponding sample size simulations (Figs. 6, 7) where the sub-sample means were distributed, on average, close to the true mean. Even with smaller sub-sample sizes, the median scarcely deviated from the true mean. Nevertheless, the results of this sample size simulations showed that under- and overestimations of the true mean are also likely in a minor number of cases.

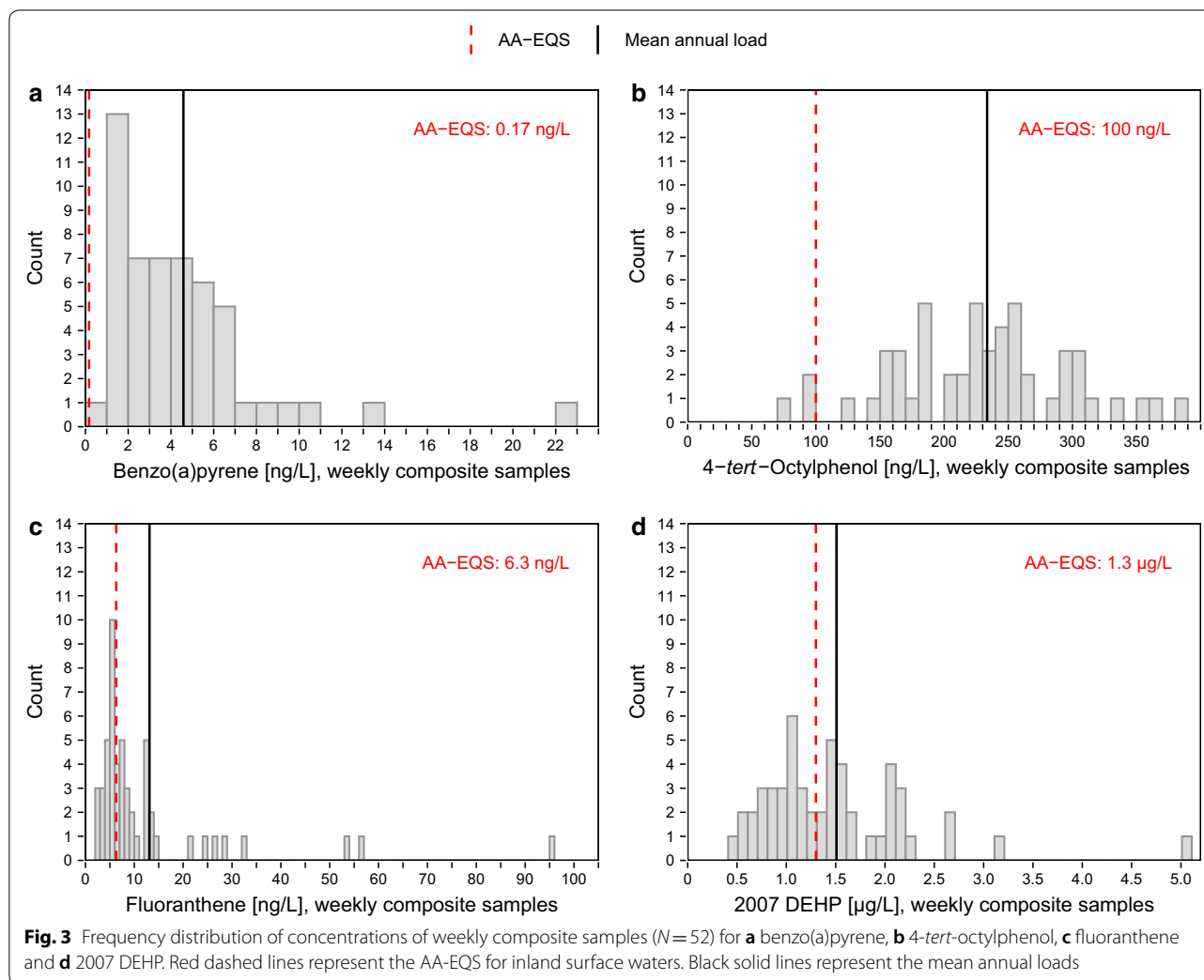
These results, overall, revealed that data set distribution was reflected in the sample size simulations and showed that the simulations differed strongly from each other due to the skewness of their original data sets. Since the data set distribution is directly reflected in the sample size simulation, it can be concluded that the observed average underestimation in smaller sub-sample sizes of right-skewed data sets is a systematically occurring error.

Sample size and temporal variability

The mean annual loads and temporal variability of DEHP data sets varied considerably and decreased between 2007 and 2016 (see Fig. 2). The results (Fig. 7) showed that the different temporal variabilities of the data sets were reflected in the corresponding sample size simulations. In detail, it could be observed that the smaller the range of the data sets, the smaller the range of the sample size simulations resulted and that the range differences were more pronounced in smaller sub-sample sizes. Therefore, it can be concluded that the outcome of sample size simulations is influenced by the skewness as well as by the temporal variability of a data set.

Influence of sample size on EQS assessment

Figure 8 depicts the proportion of false negatives for AA-EQS exceedance from the sample size simulations of fluoranthene, 2007 DEHP and 4-tert-octylphenol (see Figs. 5a, 6 and 7). False negatives for AA-EQS exceedance were defined as the proportion of the simulated sub-sample means that did not exceed the AA-EQS although the true mean exceeded the AA-EQS. Since all simulated sub-sample means of benzo(a)pyrene were above AA-EQS, no false negatives existed. For fluoranthene, 2007 DEHP and 4-tert-octylphenol, the proportion of false negatives demonstrated that the correct assessment of AA-EQS exceedance increased with increasing sub-sample size and in the case of fluoranthene and 2007 DEHP that for smaller sub-sample sizes a considerable proportion did not detect AA-EQS exceedance (Fig. 8). For example, with a sub-sample size of $n = 1$, a total of 40.4% of the fluoranthene sub-sample values and 47.9% of the DEHP sub-sample values falsely estimated the AA-EQS exceedance. These observed proportions of false negatives matched almost completely with the proportions of the 52 weekly composite samples that were below or



equal to AA-EQS (see Fig. 3). For fluoranthene, it was 40.4% (21 values) and in the case of DEHP it was 48.1% (25 values). The slight deviation of the percentage values for DEHP showed that the simulation with 100,000 runs reflected the distribution of the data sets very accurately but not completely.

The results, overall, revealed that the correct estimation of AA-EQS exceedance in the sample size simulations is influenced by the sample size as well as by the proportion of data set values below or equal to the AA-EQS.

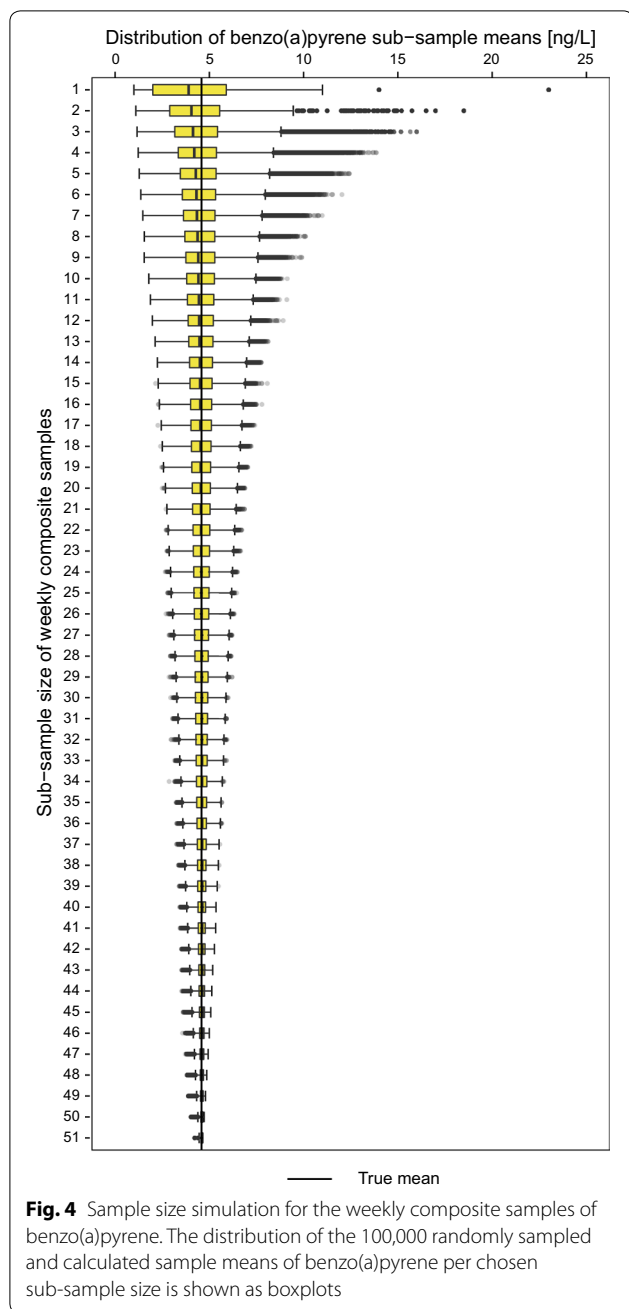
Discussion

Accuracy of the mean annual load estimation

In this study, sample size simulations were performed for priority substances, represented by benzo(a)pyrene, 4-*tert*-octylphenol, fluoranthene and DEHP. The results revealed that the accuracy of the mean annual load estimation depends very much on the sample size and the characteristics of the data sets, such as skewness and

temporal variability. Weekly composite samples provided a detailed insight into the annual pollutant concentration curves due to their high resolution. The analyzed data sets showed that they and, thus, also their sample size simulations differed greatly from each other. These differences were found between sampled years, monitoring stations and, of course, between the individual substances. It is to be expected that the differences would be even greater if the sample size simulations were performed with highly frequently measured grab samples instead of weekly composite samples. The extent of the differences, which was observed with just four substances, highlights the complexity of deriving monitoring strategies. Therefore, it is no wonder that there exists no universal monitoring strategy for (micro)pollutants as yet [47].

Regardless of how much the investigated data sets differed from each other, two patterns in the sample size simulations remained generally valid. (1) The accuracy



of the mean annual load estimation increased as the sub-sample size increased. From a purely mathematical point of view, this is not surprising. This phenomenon has also been observed for physicochemical parameters in various sampling frequency studies [48–50]. (2) Data set characteristics, such as skewness and temporal variability, were particularly reflected and more pronounced in smaller sub-sample sizes. On the basis of these two patterns, it can be concluded that the less frequent a priority substance is measured over the year and the more

variable its concentrations are over the course of the year, the more likely it is to incorrectly estimate its true mean annual load.

Small sample sizes—systematic error in the estimation of the true mean annual load in right-skewed data sets

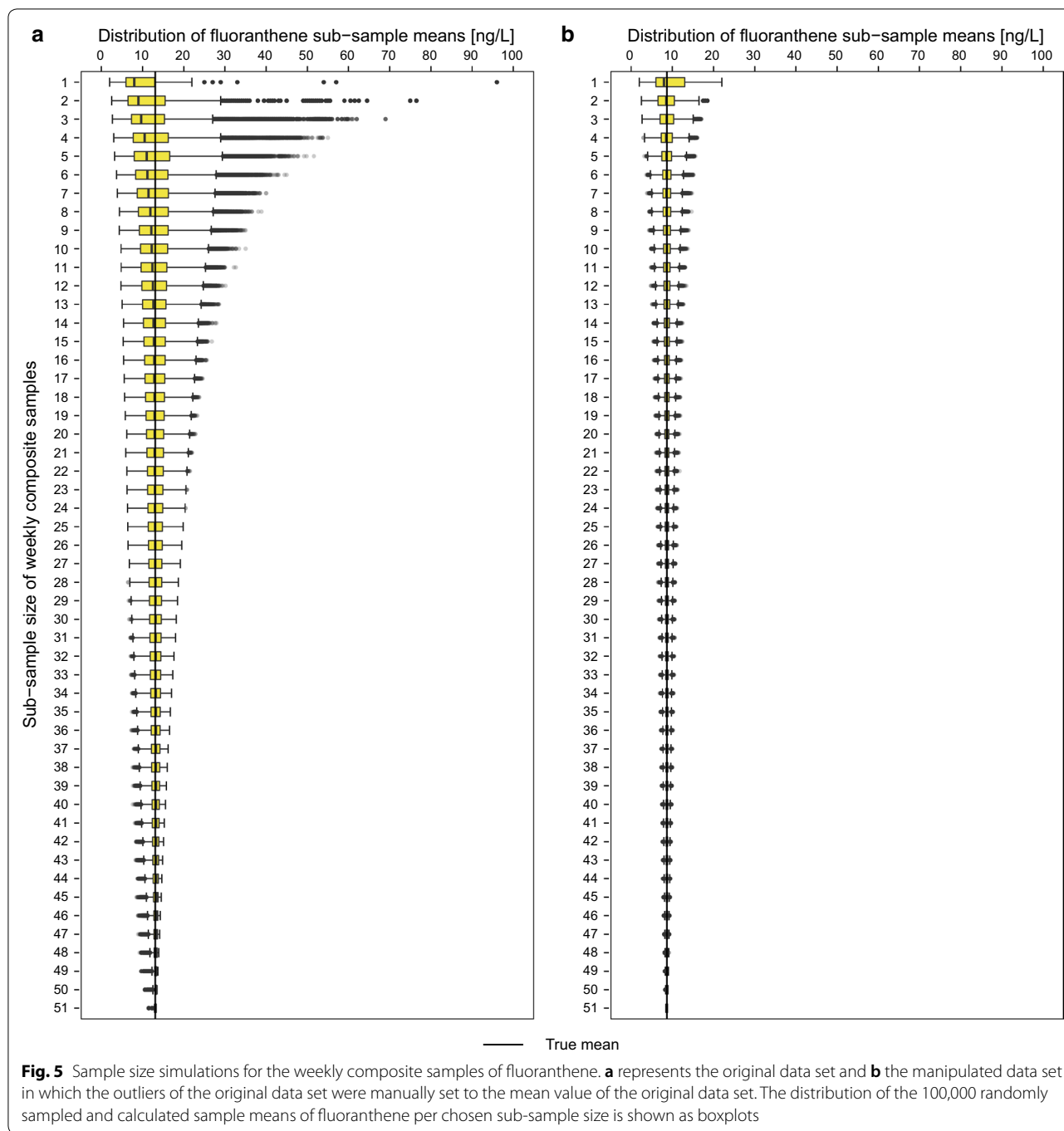
In addition to these general findings, this study was able to show that small sizes lead, on average, to a systematic underestimation of the true mean annual load in right-skewed data sets. Since the observed pattern is primarily due to data set characteristics, these findings are not only relevant for priority substances but can also be applied to other substances that occur in surface waters in a right-skewed distributed manner. These include other (micro) pollutants as well as other environmental variables, such as general physicochemical parameters. A study examining uncertainties in the annual phosphorus load estimation in rivers showed similar results and, thus, supports the stated conclusion [51].

Although the usage of small sample sizes in right-skewed data sets leads, on average, to an underestimation of the true mean annual pollutant load, the study results showed that in some cases a very strong overestimation is also possible. This means that strong uncertainties arise when small sample sizes are used because it cannot be determined afterwards whether the calculated mean annual loads are representative or whether the true mean annual loads are under- or overestimated.

Implications for monitoring programs and the implementation of the WFD

This systematic bias of under- and overestimation of the true mean annual pollutant load might also be relevant for surveillance monitoring as well as for operational monitoring. As part of operational monitoring, for example, water managers have the task of evaluating the efficiency of implemented water management measures [7]. This includes, for example, measures to reduce micropollutant inputs into surface waters, such as the upgrading or decommissioning of wastewater treatment plants, the designation of riparian buffer strips and implementation of constructed wetlands. Small sample sizes will, in these instances, lead to large uncertainties in the control results and, thus, probably compromise the validity of the efficiency control.

In addition, for the sustainable protection of freshwater ecosystems, it is essential that full compliance of the legal pollutant thresholds is maintained. For this reason, the objective must be to correctly estimate the concentrations of priority substances in surface waters occurring over the course of the year and, accordingly, to detect an annual average exceedance of the AA-EQS. As priority substances are often measured less than 12 times a



year by the European Member States for the purpose of surveillance [9] and as (micro)pollutants occur usually right-skewed distributed in surface waters [11, 20, 21], the study results suggest that the average pollutant loads of priority substances are underestimated to a greater extent throughout Europe than anticipated so far. The results of this study have also shown that the proportion of data set values below or equal to the AA-EQS basically

determines the statistical probability that EQS exceedances are not detected and that this proportion of false negatives can be very high, especially when small sample sizes are used. For the surveillance of priority substances in surface waters, it can, therefore, be concluded that the choice of sample size plays a significant role for the accuracy of EQS assessment. If small sample sizes are selected for surface water surveillance, then there is

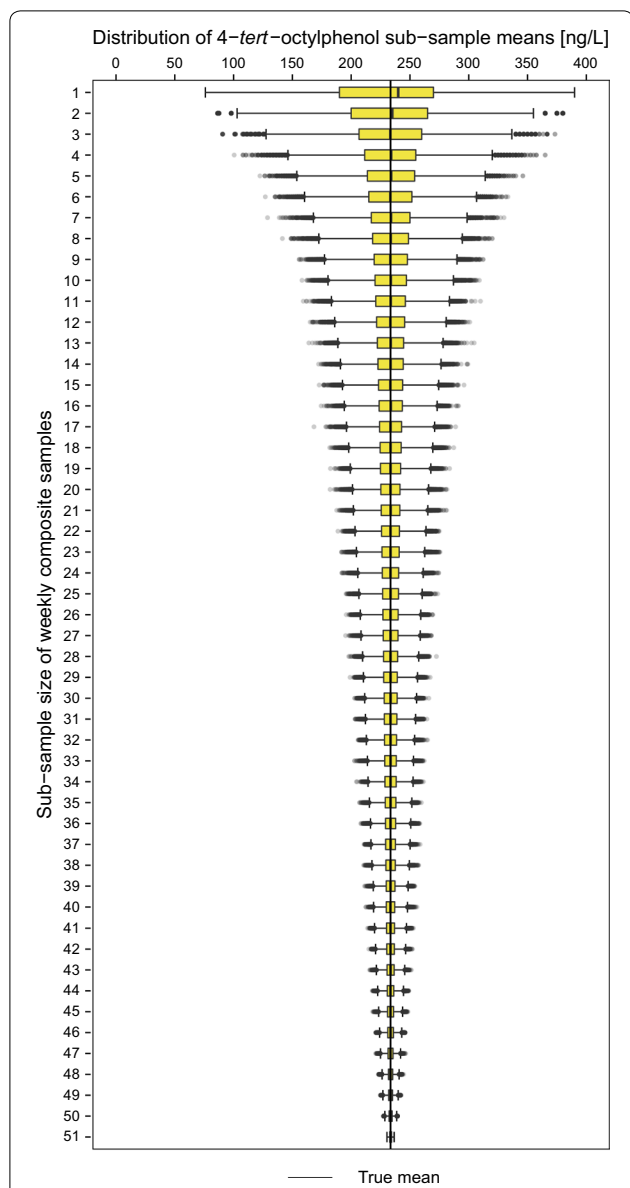


Fig. 6 Sample size simulation for the weekly composite samples of 4-tert-octylphenol. The distribution of the 100,000 randomly sampled and calculated sample means of 4-tert-octylphenol per chosen sub-sample size is shown as boxplots

an increased probability that AA-EQS exceedances will not be detected. Since, as already mentioned, sampling frequencies of less than 12 times per year are common throughout Europe, it is likely that the ecotoxicological impact of many priority substances on surface water ecosystems across Europe is underestimated. In other words, a much larger proportion of surface waters than the 46% identified in the European overview of the second River Basin Management Plans [9] may not achieve good chemical status.

Furthermore, it should be borne in mind that, in addition to the 45 priority substances, numerous other (micro)pollutants, often occurring simultaneously, pollute surface waters [52]. Therefore, a large-scale underestimation of the mean annual loads of priority substances and other (micro)pollutants could have a significant negative impact on the improvement potential of the ecological status of European surface waters. An underestimation of pollutant loads would, amongst other factors, provide an explanatory approach as to why, 18 years after the introduction of the WFD, only 40% of European surface waters have achieved good or very good ecological status and why the overall ecological status has scarcely improved between the first and second River Basin Management Plans [1]. By avoiding small sample sizes in the water surveillance of priority substances, the likely underestimated ecotoxicological risk could be better assessed and, as a consequence, be more effectively reduced by increased water management measures. These measures are expected not only to have a positive impact on chemical status but might also have a positive effect on the ecological status of surface waters in the long term.

From the results of this study, it can be concluded for monitoring programs and the implementation of the WFD that it is in principle advisable to sample more frequently to avoid uncertainties in the estimation of mean annual pollutant loads. Due to the reasons of cost efficiency, capacity efficiency and time efficiency, however, in many monitoring programs or studies only a few measurements per year are possible. In such cases, additional measures, such as the use of passive samplers [53] and event-based sampling [54], could provide solutions to assess more effectively the uncertainties.

Implications for scientific research

The findings of this study are relevant for all studies investigating the effects of environmental variables on abundance data and species composition of biocenoses by using mean annual concentrations. However, there are only very few studies that investigate the relationship between micropollutants and biological data [5, 55–58]. For one thing, this may be due to the fact that micropollutants have only been the focus of research for a few years or, for another, it may be due to the fact that, so far, too few data have been available for large-scale evaluation. Berger et al. [5] were the first to derive taxon-specific change points (for benthic invertebrates) for 6 priority substances and 19 other micropollutants using Threshold Indicator Taxa Analysis (TITAN). Change points are defined as the pollutant concentrations above which the number of individuals and the occurrence frequency of taxa abruptly decrease. The derived change

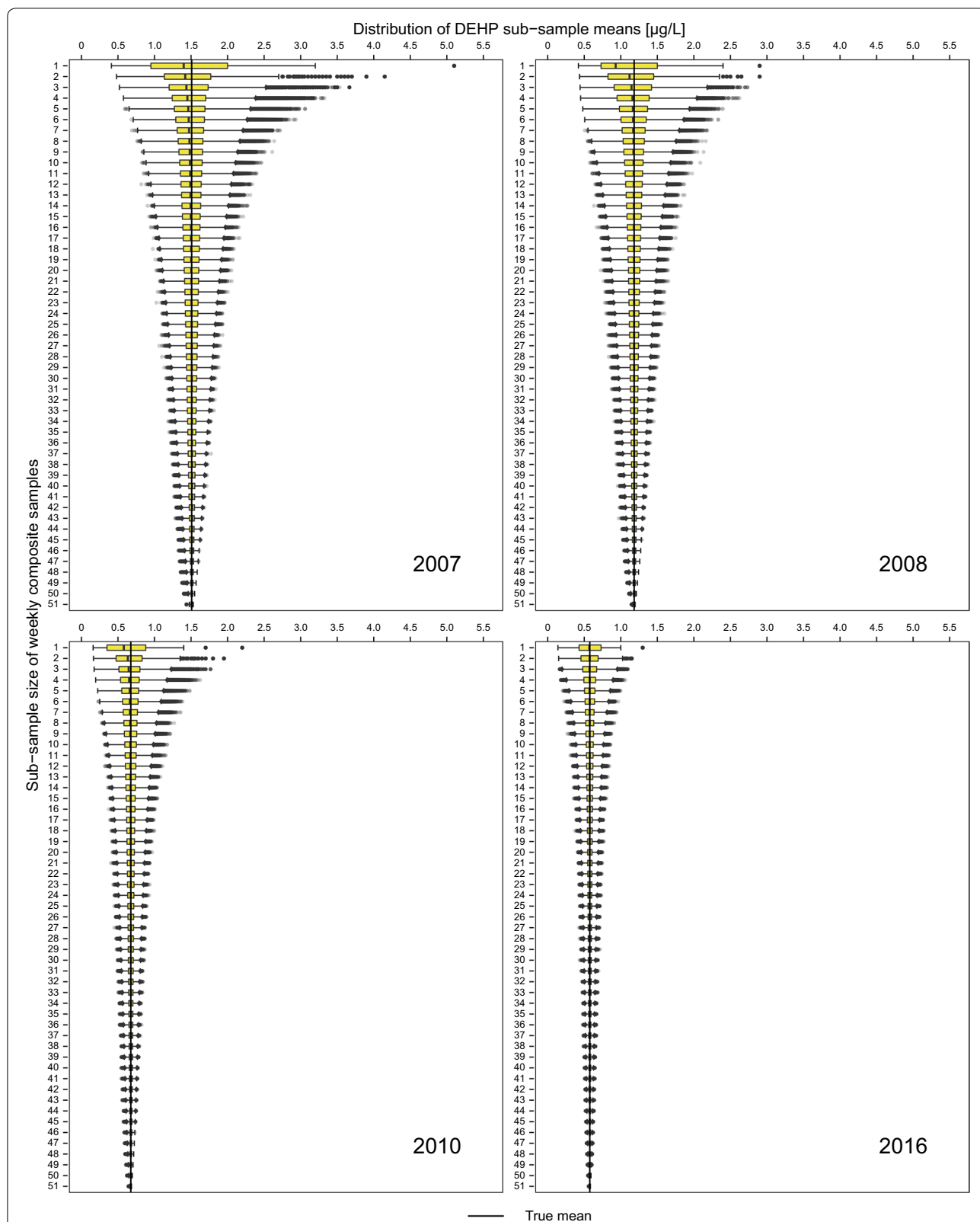
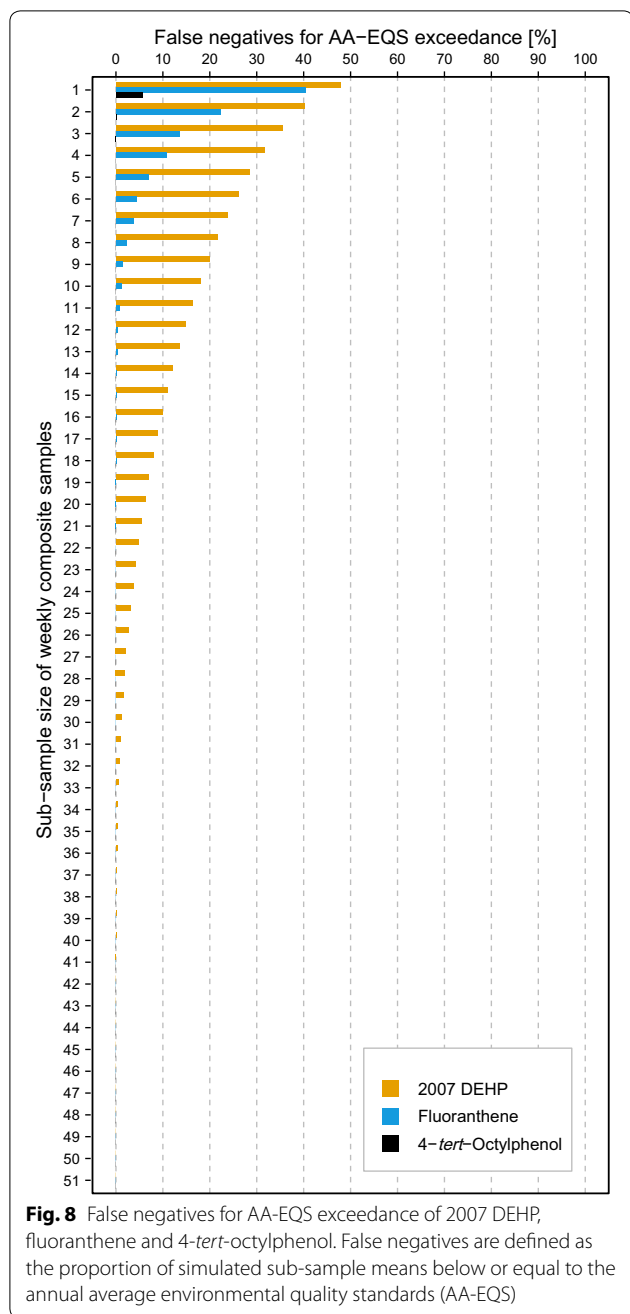


Fig. 7 Sample size simulations for the weekly composite samples of DEHP in 2007, 2008, 2010 and 2016. For each year, the distribution of the 100,000 randomly sampled and calculated sub-sample means of DEHP per chosen sub-sample size is shown as boxplots



points for many of the pollutants investigated were in comparatively low ranges and mostly far below the AA-EQS or the corresponding predicted no effect concentrations (PNECs). These results were novel and surprising from a scientific point of view and raise the critical question of whether EQS or PNECs are protective enough. Therefore, it is important to evaluate the validity of the derived change points. In simple terms, the mean annual pollutant load for each sampling site is linked with the invertebrate taxa present at that site. If only a few

pollutant values are available for the calculation of the mean, it can be expected from the results of this study that the calculated mean underestimates the true mean annual pollutant load. Under the assumption of using the true mean pollutant loads, the negative effects and the abrupt decrease of the taxa abundances compared to the calculated values should only be observable at higher pollutant concentrations. The hypothesis is, therefore, proposed that the use of small sample sizes from right-skewed and temporally variable pollutant concentrations results in the derived change points being too low. Since Berger et al. [5] included micropollutants with a sampling frequency of at least 4 values per year for the calculation of the mean annual pollutant loads, it would be quite possible that the calculated change points are too low. However, it remains to be clarified whether and to what extent small sample sizes of micropollutant concentrations affect the derived change points.

In multivariate analyses, it is investigated which environmental parameters form freshwater biocenoses and act as stressors on them [59]. In these so-called multiple stressor studies, very diverse environmental parameters are used in the models and linked to the biocenoses. Some environmental parameters, such as altitude and geology, are static. Other parameters have a low temporal variability, such as land use, land cover and distance to the closest wastewater treatment plant, or a high temporal variability, such as many physicochemical parameters and, of course, (micro)pollutants. In the case of parameters that are static or show only slight temporal variability, it is sufficient to determine these parameters once per study period or once per year, respectively, in order to be able to reflect them representatively. However, as the results of this study have shown, this does not apply to parameters with a high temporal variability, as their true annual concentration curves are better represented the more often that they are measured in the course of a year. For this reason, it is hypothesized that in the case of environmental concentrations with high temporal variability, individually measured values or mean values determined from small sample sizes diminish the model performance of multivariate analyses. Since monitoring data is often scarce, especially for large-scale studies, many studies have to work with small sample sizes [55–58]. In such studies, scientists should bear in mind that the explanatory power of the relationships between environmental parameters with high temporal variability and freshwater biocenoses is likely to be reduced due to the small sample size.

Conclusions

This study showed that the choice of sample size plays a significant role in the monitoring of priority substances in surface waters and that water managers and scientists need to be aware that small sample sizes can lead to large uncertainties and potentially jeopardize the validity of their research results. It was revealed that a small sample size is likely to result in the underestimation of true mean annual pollutant loads and in the failure to detect EQS exceedances. Even though this study did not aim to optimize the sampling frequency, the results emphasize the importance of sampling priority substances according to the WFD guidelines at least 12 times a year to ensure the protection of the freshwater ecosystems in Europe.

Abbreviations

AA-EQS: annual average environmental quality standard; DEHP: di(2-ethylhexyl) phthalate; EQS: environmental quality standard; PAHs: polycyclic aromatic hydrocarbons; PNEC: predicted no effect concentration; SD: standard deviation; TITAN: Threshold Indicator Taxa Analysis; WFD: European Water Framework Directive.

Acknowledgements

The first author gratefully acknowledges the financial support received towards her Ph.D. from the Hans Böckler Foundation Ph.D. fellowship. The authors would like to thank Jörg Oehlmann for his critical comments on an earlier version of the manuscript. Particular thanks go out to the Saxon State Office for Environment, Agriculture and Geology for the provision of micropollutant data (<https://www.umwelt.sachsen.de/umwelt/wasser/7112.htm>). English Language editing service was supplied by Goethe Research Academy for Early Career Researchers.

Authors' contributions

DB and AS conceived and designed the study. DB acquired the data and performed the study. DB and AS wrote the article. Both authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The micropollutant data sets used in the present study are publicly available from the Saxon State Office for Environment, Agriculture and Geology at <https://www.umwelt.sachsen.de/umwelt/wasser/7112.htm>.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹ Department of River Ecology and Conservation, Senckenberg Research Institute and Natural History Museum Frankfurt, Clameystr. 12, 63571 Gelnhausen, Germany. ² Institute of Ecology, Evolution and Diversity, Department Aquatic Ecotoxicology, Goethe University, Max-von-Laue-Str. 13, 60438 Frankfurt am Main, Germany.

Received: 9 September 2019 Accepted: 29 December 2019
Published online: 16 January 2020

References

- European Environment Agency (2018) European waters—Assessment of status and pressures 2018. European Environment Agency, Copenhagen. <https://doi.org/10.2800/303664>
- Schäfer RB, Kühn B, Malaj E, König A, Gergs R (2016) Contribution of organic toxicants to multiple stress in river ecosystems. *Freshw Biol* 61:2116–2128. <https://doi.org/10.1111/fwb.12811>
- Malaj E, von der Ohe PC, Grote M, Kühne R, Mondy CP, Usseglio-Polatera P, Brack W, Schäfer RB (2014) Organic chemicals jeopardize the health of freshwater ecosystems on the continental scale. *Proc Natl Acad Sci* 111:9549–9554. <https://doi.org/10.1073/pnas.1321082111>
- Stehle S, Schulz R (2015) Agricultural insecticides threaten surface waters at the global scale. *Proc Natl Acad Sci* 112:5750–5755. <https://doi.org/10.1073/pnas.1500232112>
- Berger E, Haase P, Oetken M, Sundermann A (2016) Field data reveal low critical chemical concentrations for river benthic invertebrates. *Sci Total Environ* 544:864–873. <https://doi.org/10.1016/j.scitotenv.2015.12.006>
- Münze R, Hannemann C, Orlinskiy P, Gunold R, Paschke A, Foit K, Becker J, Kaske O, Paulsson E, Peterson M, Jernstedt H, Kreuger J, Schüürmann G, Liess M (2017) Pesticides from wastewater treatment plant effluents affect invertebrate communities. *Sci Total Environ* 599–600:387–399. <https://doi.org/10.1016/j.scitotenv.2017.03.008>
- European Parliament and Council (2000) Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy, L327. European Parliament and Council, Brussels
- European Parliament and Council (2013) Directive 2013/39/EU of the European Parliament and of the Council of 12 August 2013 amending Directives 2000/60/EC and 2008/105/EC as regards priority substances in the field of water policy, L 226. European Parliament and Council, Brussels
- European Commission (2019) Commission staff working document. European Overview—river basin management plans. Accompanying the document 'Report from the Commission to the European Parliament and the Council implementation of the Water Framework Directive (2000/60/EC) and the Floods Directive (2007/60/EC), Second River Basin Management Plans, First Flood Risk Management Plans'
- Sousa JCG, Ribeiro AR, Barbosa MO, Pereira FR, Silva AMT (2018) A review on environmental monitoring of water organic pollutants identified by EU guidelines. *J Hazard Mater* 344:146–162. <https://doi.org/10.1016/j.jhazmat.2017.09.058>
- Vryzas Z, Vassiliou G, Alexoudis C, Papadopoulou-Mourkidou E (2009) Spatial and temporal distribution of pesticide residues in surface waters in northeastern Greece. *Water Res* 43:1–10. <https://doi.org/10.1016/j.watres.2008.09.021>
- Chon H-S, Ohandja D-G, Voulvoulis N (2010) Implementation of E.U. Water framework directive: source assessment of metallic substances at catchment levels. *J Environ Monit* 12:36–47. <https://doi.org/10.1039/b907851g>
- Christoffels E, Brunsch A, Wunderlich-Pfeiffer J, Mertens FM (2016) Monitoring micropollutants in the Swist river basin. *Water Sci Technol* 74:2280–2296. <https://doi.org/10.2166/wst.2016.392>
- Petrucci G, Gromaire M-C, Shorshani MF, Ghebbo G (2014) Nonpoint source pollution of urban stormwater runoff: a methodology for source analysis. *Environ Sci Pollut Res* 21:10225–10242. <https://doi.org/10.1007/s11356-014-2845-4>
- Gerrity D, Trenholm RA, Snyder SA (2011) Temporal variability of pharmaceuticals and illicit drugs in wastewater and the effects of a major sporting event. *Water Res* 45:5399–5411. <https://doi.org/10.1016/j.watres.2011.07.020>
- Nelson ED, Do H, Lewis RS, Carr SA (2011) Diurnal variability of pharmaceutical, personal care product, estrogen and alkylphenol concentrations in effluent from a tertiary wastewater treatment facility. *Environ Sci Technol* 45:1228–1234. <https://doi.org/10.1021/es102452f>
- Mandarić L, Diamantini E, Stella E, Cano-Paoli K, Valle-Sistac J, Molins-Delgado D, Bellin A, Chiogna G, Majone B, Diaz-Cruz MS, Sabater S, Barceló D, Petrovic M (2017) Contamination sources and distribution patterns of pharmaceuticals and personal care products in Alpine rivers strongly affected by tourism. *Sci Total Environ* 590–591:484–494. <https://doi.org/10.1016/j.scitotenv.2017.02.185>
- Musolff A, Leschik S, Möder M, Strauch G, Reinstorf F, Schirmer M (2009) Temporal and spatial patterns of micropollutants in urban receiving

- waters. *Environ Pollut* 157:3069–3077. <https://doi.org/10.1016/j.envpol.2009.05.037>
19. Osorio V, Marcé R, Pérez S, Ginebreda A, Cortina JL, Barceló D (2012) Occurrence and modeling of pharmaceuticals on a sewage-impacted Mediterranean river and their dynamics under different hydrological conditions. *Sci Total Environ* 440:3–13. <https://doi.org/10.1016/j.scitotenv.2012.08.040>
 20. Gardner MJ (2014) Lognormality of trace contaminant concentrations in sewage effluents. *Environ Monit Assess* 186:4819–4827. <https://doi.org/10.1007/s10661-014-3740-7>
 21. Ott WR (1990) A physical explanation of the lognormality of pollutant concentrations. *J Air Waste Manag Assoc* 40:1378–1383. <https://doi.org/10.1080/10473289.1990.10466789>
 22. Doppler T, Lück A, Camenzuli L, Krauss M, Stamm C (2014) Critical source areas for herbicides can change location depending on rain events. *Agric Ecosyst Environ* 192:85–94. <https://doi.org/10.1016/j.agee.2014.04.003>
 23. Zgheib S, Moilleron R, Chebbo G (2012) Priority pollutants in urban stormwater: part 1—case of separate storm sewers. *Water Res* 46:6683–6692. <https://doi.org/10.1016/j.watres.2011.12.012>
 24. Gasperi J, Zgheib S, Cladière M, Rocher V, Moilleron R, Chebbo G (2012) Priority pollutants in urban stormwater: part 2—case of combined sewers. *Water Res* 46:6693–6703. <https://doi.org/10.1016/j.watres.2011.09.041>
 25. Launay MA, Dittmer U, Steinmetz H (2016) Organic micropollutants discharged by combined sewer overflows—characterisation of pollutant sources and stormwater-related processes. *Water Res* 104:82–92. <https://doi.org/10.1016/j.watres.2016.07.068>
 26. Weyrauch P, Matzinger A, Pawlowsky-Reusing E, Plume S, von Seggern D, Heinzmann B, Schroeder K, Rouault P (2010) Contribution of combined sewer overflows to trace contaminant loads in urban streams. *Water Res* 44:4451–4462. <https://doi.org/10.1016/j.watres.2010.06.011>
 27. Bundschuh M, Zubrod JP, Klemm P, Elsaesser D, Stang C, Schulz R (2013) Effects of peak exposure scenarios on *Gammarus fossarum* using field relevant pesticide mixtures. *Ecotoxicol Environ Saf* 95:137–143. <https://doi.org/10.1016/j.ecoenv.2013.05.025>
 28. Zhao X-M, Yao L-A, Ma Q-L, Zhou G-J, Wang L, Fang Q-L, Xu Z-C (2018) Distribution and ecological risk assessment of cadmium in water and sediment in Longjiang River, China: implication on water quality management after pollution accident. *Chemosphere* 194:107–116. <https://doi.org/10.1016/j.chemosphere.2017.11.127>
 29. Thompson SK (2012) Sampling, 3rd edn. Wiley, Hoboken
 30. Choi S-D (2014) Time trends in the levels and patterns of polycyclic aromatic hydrocarbons (PAHs) in pine bark, litter, and soil after a forest fire. *Sci Total Environ* 470–471:1441–1449. <https://doi.org/10.1016/j.scitotenv.2013.07.100>
 31. Vergnoux A, Malleret L, Asia L, Doumenq P, Theraulaz F (2011) Impact of forest fires on PAH level and distribution in soils. *Environ Res* 111:193–198. <https://doi.org/10.1016/j.envres.2010.01.008>
 32. Stracquadanio M, Dinelli E, Trombini C (2003) Role of volcanic dust in the atmospheric transport and deposition of polycyclic aromatic hydrocarbons and mercury. *J Environ Monit* 5:984–988. <https://doi.org/10.1039/b308587b>
 33. Kozielska B, Koniecznyński J (2015) Polycyclic aromatic hydrocarbons in particulate matter emitted from coke oven battery. *Fuel* 144:327–334. <https://doi.org/10.1016/j.fuel.2014.12.069>
 34. Liberti L, Notarnicola M, Primerano R, Zannetti P (2006) Air pollution from a large steel factory: polycyclic aromatic hydrocarbon emissions from coke-oven batteries. *J Air Waste Manag Assoc* 56:255–260. <https://doi.org/10.1080/10473289.2006.10464461>
 35. Lima ALC, Farrington JW, Reddy CM (2005) Combustion-derived polycyclic aromatic hydrocarbons in the environment—a review. *Environ Forensics* 6:109–131. <https://doi.org/10.1080/15275920590952739>
 36. Napier F, D'Arcy B, Jefferies C (2008) A review of vehicle related metals and polycyclic aromatic hydrocarbons in the UK environment. *Desalination* 226:143–150. <https://doi.org/10.1016/j.desal.2007.02.104>
 37. Baumann W, Ismeier M (1998) Natural rubber and rubber: Facts and figures on environmental protection (Kautschuk und Gummi: Daten und Fakten zum Umweltschutz), vol 1–2. Springer, Berlin
 38. Wagner BO, Mücke W, Schenck H-P (1989) Environmental monitoring: Environmental concentrations of organic chemicals—literature research and evaluation (Umwelt-Monitoring: Umweltkonzentrationen organischer Chemikalien—Literatur-Recherche und -Auswertung). Ecomed Verlagsgesellschaft mbH, Landsberg am Lech
 39. Baumann W, Herberg-Liedtke B (1996) Chemicals in metal processing—facts and figures on environmental protection (Chemikalien in der Metallbearbeitung—Daten und Fakten zum Umweltschutz). Springer, Berlin. <https://doi.org/10.1007/978-3-642-61004-2>
 40. Brooke D, Johnson I, Mitchell R, Watts C (2005) Environmental risk evaluation report: 4-tert-octylphenol. Environment Agency, Bristol
 41. Fuchs S, Rothvoß S, Toshovski S (2018) Ubiquitous pollutants—Entry path inventories, environmental behaviour and entry path modelling (Ubiquitäre Schadstoffe—Eintragsinventare, Umweltverhalten und Eintragsmodellierung. Forschungsbericht 21 200 0 UBA-FB 002648). Research Report 3714 21 200 0 UBA-FB 002648. Federal Environment Agency, Dessau-Rosslau
 42. Joint Research Center (2008) Bis (2-ethylhexyl) phthalate (DEHP) Summary Risk Assessment Report
 43. European Chemicals Agency (2008) Inclusion of substances of very high concern in the candidate list (Decision by the Executive Director). European Chemicals Agency, Helsinki
 44. Commission Europe (2011) Commission regulation (EU) No 143/2011 of 17 February 2011 amending Annex XIV to regulation (EC) No 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals ('REACH'), L44. European Commission, Brussels
 45. Tukey JW (1977) Exploratory data analysis. Addison-Wesley Publishing Company, Reading
 46. R Core Team (2018) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.r-project.org/>
 47. Behmel S, Damour M, Ludwig R, Rodriguez MJ (2016) Water quality monitoring strategies—a review and future perspectives. *Sci Total Environ* 571:1312–1329. <https://doi.org/10.1016/j.scitotenv.2016.06.235>
 48. Birgand F, Faucheux C, Gruau G, Augeard B, Moatar F, Bordenave P (2010) Uncertainties in assessing annual nitrate loads and concentration indicators. Part 1: impact of sampling frequency and load estimation algorithms. *Trans Am Soc Agric Biol Eng* 53:437–446
 49. Skeffington RA, Halliday SJ, Wade AJ, Bowes MJ, Loewenthal M (2015) Using high-frequency water quality data to assess sampling strategies for the EU Water Framework Directive. *Hydro Earth Syst Sci* 19:2491–2504. <https://doi.org/10.5194/hess-19-2491-2015>
 50. Valkama P, Ruth O (2017) Impact of calculation method, sampling frequency and Hysteresis on suspended solids and total phosphorus load estimations in cold climate. *Hydro Res* 48:1594–1610. <https://doi.org/10.2166/nh.2017.199>
 51. Johns PJ (2007) Uncertainties in annual riverine phosphorus load estimation: impact of load estimation methodology, sampling frequency, baseflow index and catchment population density. *J Hydrol* 332:241–258. <https://doi.org/10.1016/j.jhydrol.2006.07.006>
 52. Schwarzenbach RP, Escher BI, Fenner K, Hofstetter TB, Johnson CA, von Gunten U, Wehrli B (2006) The challenge of micropollutants in aquatic systems. *Science* 313:1072–1077. <https://doi.org/10.1126/science.1127291>
 53. Lorenz S, Rasmussen JJ, Süß A, Kalettka T, Golla B, Horney P, Stähler M, Hommel B, Schäfer RB (2017) Specifics and challenges of assessing exposure and effects of pesticides in small water bodies. *Hydrobiologia* 793:213–224. <https://doi.org/10.1007/s10750-016-2973-6>
 54. Stehle S, Knäbel A, Schulz R (2013) Probabilistic risk assessment of insecticide concentrations in agricultural surface waters: a critical appraisal. *Environ Monit Assess* 185:6295–6310. <https://doi.org/10.1007/s10661-012-3026-x>
 55. Giulivo M, Stella E, Capri E, Esnaola A, López de Alda M, Diaz-Cruz S, Mandaric L, Muñoz I, Bellin A (2019) Assessing the effects of hydrological and chemical stressors on macroinvertebrate community in an Alpine river: the Adige River as a case study. *River Res Appl* 35:78–87. <https://doi.org/10.1002/rra.3367>
 56. Muñoz I, López-Doval J, Ricart M, Villagrasa M, Brix R, Geiszinger A, Ginebreda A, Guasch H, López de Alda M, Romani A, Sabater S, Barceló D (2009) Bridging levels of pharmaceuticals in river water with biological community structure in the Llobregat river basin (northeast Spain). *Environ Toxicol Chem* 28:2706–2714. <https://doi.org/10.1897/08-486.1>

57. Sabater S, Barceló D, De Castro-Català N, Ginebreda A, Kuzmanovic M, Petrovic M, Picó Y, Ponsatí L, Tornés E, Muñoz I (2016) Shared effects of organic microcontaminants and environmental stressors on biofilms and invertebrates in impaired rivers. *Environ Pollut* 210:303–314. <https://doi.org/10.1016/j.envpol.2016.01.037>
58. Smeti E, von Schiller D, Karaouzas I, Laschou S, Vardakas L, Sabater S, Tornés E, Monllor-Alcaraz LS, Guillem-Argiles N, Martínez E, Barceló D, López de Alda M, Kalogianni E, Elosegi A, Skoulikidis N (2019) Multiple stressor effects on biodiversity and ecosystem functioning in a Mediterranean temporary river. *Sci Total Environ* 647:1179–1187. <https://doi.org/10.1016/j.scitotenv.2018.08.105>
59. Hernandez-Suarez S, Nejadhashemi AP (2018) A review of macroinvertebrate- and fish-based stream health modelling techniques. *Ecohydrology* 11:1–24. <https://doi.org/10.1002/eco.2022>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)
