Environmental Sciences Europe

RESEARCH

Open Access

A harmonized Danube basin-wide multi-compartment concentration database to support inventories of micropollutant emissions to surface waters

Steffen Kittlaus[1*], Máté Krisztián Kardos[2], Katalin Mária Dudás[2], Nikolaus Weber[1], Adrienne Clement[2], Silviya Petkova[3], Danijela Sukovic[4], Dajana Kučić Grgić[5], Adam Kovacs[6], David Kocman[7], Constanta Moldovan[8], Michal Kirchner[9], Oliver Gabriel[10], Jörg Krampe[1], Matthias Zessner[1] and Ottavia Zoboli[1]

Abstract

Background The European Water Framework Directive foresees the establishment of emission inventories for micropollutants (MP) to facilitate an evidence-based development of mitigation measures. Regionalized pathway analysis constitutes a moderately data-intensive approach to quantify the contribution of different pathways to the total pollution of surface waters. So far, only few European member states have created an inventory that includes diffuse pathways. The fundamental basis to enable it is an accessible, well-structured and harmonized database with data on the concentration of MPs in multiple compartments, such as soils, groundwater, atmospheric deposition and urban systems. Combined with the water and suspended substance balance in river basins, such data enables the estimation of emission loads via specific pathways. In the Danube River Basin, but in general in Europe, a public data management platform with such scope and criteria is still lacking.

Results We collected and harmonized MP measurements across multiple compartments and countries together with key metadata, harmonized and combined them into a new database. The resulting tool, available for download, facilitates the assessment of current data availability, in terms of quantity and quality. For example, while the majority of available data stems from groundwater and surface water, other highly relevant compartments are scarcely represented. By examining differences in MP concentration level across compartments, the database can lead to understand the relevance of specific emission pathways and thus to prioritize data-retrieval and calculation efforts in modelling applications. Selected examples show how to exploit the metadata associated to the measurements to extrapolate the results to regions not covered by specific monitoring programmes. For example, PFAS concentrations in treated wastewater show significant dependence on the design capacity of the treatment plant.

Conclusions This study showcases how such database can support the setup of emission inventories, guide data providers and national authorities in prioritizing the allocation of resources for new surveys and in optimizing their national data collection and management systems. The process tested showed a great need for enhanced data

*Correspondence:
Steffen Kittlaus
steffen.kittlaus@tuwien.ac.at
Full list of author information is available at the end of the article

literacy across countries and institutions to increase data availability and quality to secure the exploitation of the full information potential generated via monitoring programmes.

Keywords Trace contaminants, Emission inventory, Surface water pollution, Concentration database, Regionalized pathway analysis, Wastewater, Groundwater, Atmospheric deposition, Soil, Stormwater runoff

Background

The European Water Framework Directive [1] and its daughter directives define a policy regime to improve the quality of surface waters and is considered one of the world's most advanced approaches in this area. Regarding micropollutants (MP), besides requirements for monitoring in surface water bodies as basis for the chemical status assessment, member states are required to set up inventories of emissions, discharges and losses of priority and priority hazardous substances to facilitate the evidence-based development of emission mitigation measures. The guidance document No. 28 [25] presents a tiered methodological approach that should be followed to establish such inventories, depending on data availability and substance-criticality. The riverine load approach (tier 2) relies on hydrological and chemical measurements in the rivers to estimate the mass of contaminants transported per unit of time. When combined with the inventory of point source emissions (tier 1) and considering in-river processes such as degradation or transformation, it allows estimation of the proportion of diffuse emissions. These are, however, treated as a black box. To provide an accurate inventory of point and specific diffuse emissions, the next level (tier 3) is necessary. This involves using a pathway-oriented approach or regionalized pathway analysis (RPA). Its application requires riverine loads from tier 1 to validate the estimated emissions, but also more information on land use, hydrology and main processes in the river catchments. In particular, data on occurrence and concentration levels of contaminants in multiple compartments, such as soils, groundwater, atmospheric deposition and urban systems are essential. In order to obtain a comprehensive depiction of the life cycle of contaminants, the source-oriented approach (tier 4) is designed to estimate emissions factors for different human activities, starting from substance-specific data on production, sales and consumption. As the tier increases, so does our understanding of sources and pathways, as well as our capability to identify the most appropriate measures. However, higher tiers are also associated with a larger amount and higher complexity of data needs.

The second cycle of river basin management plan development in 2015 revealed that only a limited number of EU member states were able to report data and estimations of diffuse emissions and only for a limited range of substances [64]. This is also the case for the Danube River Basin (DRB), which is one of the most international river basins in the world [63]. Only a handful of countries in the DRB have explicitly included diffuse emissions according to tier 3 in their inventories, while all others have only created an inventory of point sources. As of the update of the DRB management plan in 2021, no transnational emission inventory was available [39].

The first fundamental step to enable the generation and regular update of an accurate emissions inventory is to have an accessible, well-structured and harmonized data basis that contains information on the concentration of MPs in various environmental and technical compartments. This information, when combined with the water and suspended solids balance of the basin, allows for the estimation of emission loads through specific pathways. Due to the impossibility of monitoring the whole territory, it is essential to have thorough documentation and accessible metadata associated with surveys in different compartments. This will enable the identification of statistically significant patterns and correlations, allowing for the extrapolation and interpolation of data for unmonitored locations and regions.

Multiple international initiatives have been launched recently to collect data on MPs at European and DRB scale:

The NORMAN network maintains the "EMPODAT" database [62] to support the identification and prioritization of pollutants of emerging concern. It focuses on new substances that are known, but not yet well investigated [22]. In line with this purpose, the database contains several metadata on the analytical methods and analytical quality assessment. Metadata on the sampled environmental compartment are included to a lesser extent. The data cover many substances with measurements at different levels of quantitative precision, ranging from uncalibrated screening methods to fully quantitative target analysis with application of quality control procedures. In line with its purpose, the database does not collect monitoring data for well-known substances such as heavy metals or from national administrations. The absence of measurements for well-known parameters, missing metadata about the sampling sites and facilities, and semi-quantitative

measurements in the database impede its use as a basis for emission modelling. Nevertheless, the metadata regarding analytical methods can serve as a best-practice example.

The EU joint research centre (JRC) maintains the "JRC FATE Monitoring Database on Occurrence and Levels of Chemical Contaminants" [35], probably containing reported data from the EU member states. Unfortunately, this database is not public.

Many EU and non-EU countries report environmental data to the European Environmental Agency (EEA) within the EIONET network. The EEA then publishes this data, such as the concentration data for surface and groundwater bodies found in the "Waterbase—Water Quality ICM" dataset [30]. This dataset provides only limited metadata on analysis methods and sampling site properties and, as it covers only ground and surface water, it cannot be used as a basis for an inventory of multiple emission pathways. Comparable datasets for other compartments, e.g. soil or atmospheric deposition covering more than a few substances, are not yet available from the EEA.

The European Industrial Emissions Portal [29] publishes the emissions to water reported under the Pollutant Release and Transfer Register (PRTR) directive 2 by the EU member states and some non-member states, such as Serbia. The reporting of emissions is restricted to industries belonging to the 65 economic activities listed in Annex I of the directive and exceeding at least one of the PRTR capacity thresholds (regarding production or processing volumes). Additionally, only data on emissions of pollutants that exceed thresholds specified in Annex II must be released. If industries meet these preconditions, they must report emissions to water for 71 pollutants, but only where they exceed the given threshold. The reported emissions in mass per year can be based on measurements, calculations or estimations. Due to the rather high pollutant-specific thresholds, only major polluters are required to report their emissions. Further, the reported emissions are subject to large uncertainties depending on the method applied for quantification. For MPs, values below the analytical limit of quantification are often regarded as 0, resulting in a best-case evaluation of emission loads. To extrapolate the reported loads to other emitters who are not required to report their emissions for the aforementioned reasons, or to estimate emissions for substances not included in the PRTR, one would need the water volumes and concentration data used to calculate the loads. Even if the water volumes are known—which is mostly not the case, as they do not need to be reported—the aforementioned reasons make the extrapolation highly uncertain.

The International Commission for the Protection of the Danube River (ICPDR) operates the "Danube River Basin Water Quality Database" [40], which includes data from the "TransNational Monitoring Network" (TNMN) and the "Joint Danube Surveys" (JDS). The TNMN contains analyses of water samples from selected stations along the Danube for a limited number of MP. To complement this substance-wise rather limited monitoring, the JDS collects samples along the Danube once every 6 years and analyses them for a very wide range of pollutants. The content of this database focuses on contamination in the Danube River itself. The datasets only include a few major tributaries of the Danube, and do not cover other environmental compartments that contribute to river pollution, such as atmospheric deposition, stormwater or soil. Although two other compartments (wastewater and groundwater) were sampled during JDS4, their data are not yet included in the database system. To use this database as basis for basin-wide emission modelling, a massive extension of the scope would be required. This includes expanding the substance list, temporal and spatial coverage, and the covered environmental compartments.

Finally, the French newspaper "Le Monde" has compiled an extensive dataset on per- and polyfluoroalkyl substances (PFAS) pollution in Europe, which also includes data from various environmental compartments [20]. However, the amount of metadata included is different as is the purpose of this web application.

A database which shows comparable efforts outside of Europe is the one established by the U.S. Environmental Protection Agency (US-EPA). This database compiles pre-existing data to support exposure assessment and has a wider focus regarding the environmental compartments covered [41]. The data can be traced back to the original source files. However, the database in its published state is not a suitable basis to set up emission inventories due to the limited metadata regarding the sampled environmental compartments.

This brief overview clearly shows that these important and valuable initiatives were launched with a wide range of objectives, none of which included the creation of an accurate inventory of point and diffuse emissions. Therefore, they were not designed to meet the data and metadata requirements of such endeavour. A public database or data management platform with this scope and criteria is still absent in the DRB and Europe as a whole.

To address this critical gap, the Danube Hazard m^3c project [23] undertook a major effort to collect and harmonize data and metadata on MP measurements across multiple compartments and Danube countries. This was done to create a new database with the goal of providing the optimal information basis for compiling a

transnational emission inventory according to the pathway-oriented approach.

In doing so, we addressed several research questions. Firstly, we investigated the current availability of data in the different compartments and regions of the basin, considering both quantity and quality. The results of this analysis are important, as they provide modellers with a basis to assess the varying levels of uncertainty affecting the estimation of emission loads through different pathways. This, in turn, affects the uncertainty in predicting the effectiveness of measures implemented in different compartments. Additionally, the results provide data providers and national authorities with criteria to prioritize the allocation of resources in designing new surveys and optimizing their national data collection and management systems. Secondly, we examined the differences in the occurrence and concentration levels of specific substances across different compartments and countries. This screening helps us to better understand the relevance or dominance of specific emission pathways, allowing us to prioritize data-retrieval and calculation efforts in modelling applications. Moreover, one of the most important objectives was to explore the metadata to assess the potential for extrapolation of information for estimating regionalized emissions in areas of the basin that are not monitored by specific programmes.

This paper presents and critically discusses the conceptual design of the database, the lessons learned during its implementation, the criteria used to select the collected data and metadata, and the new knowledge and added value that can be obtained through their analysis. It will be highly valuable not only for institutions dealing with water quality management in the DRB, but also for scientists or public authorities interested in launching a similar endeavour in other river basins worldwide.

Material and methods

Setup of the database

The database is implemented as relational database in PostgreSQL [56], as this is a powerful and open source database management system.

The data are organized into 31 main tables which contain the actual data, and 38 supporting tables that contain the allowed entries for columns in the main tables with controlled vocabularies. The use of controlled vocabularies is of utmost importance to harmonize the metadata for data from different data sources and make it evaluable as one.

All tables include information about the data source, the date of data import, and the user responsible. The main tables containing concentration data include metadata on the analysed matrix (total, dissolved or solid phase), sample preparation and analytical methods along with their limit of quantitation (LOQ) and, where applicable, limit of detection (LOD), as well as references to national or ISO norms and the laboratory responsible for the analysis.

Separate main tables include data for the different environmental and technical compartments:

Water bodies: concentrations from river and groundwater bodies along with suspended particulate matter (SPM) from rivers. The metadata for river samples includes the sampling method and the discharge situation at the time of sampling compared to the long-term mean discharge. Similarly, for groundwater samples, the metadata includes the sampling depth and the water level at the time of sampling, compared to the long-term water level. For river monitoring sites, besides the exact location (coordinates with coordinate reference system and country), information on the catchment size and correlated other monitoring sites (e.g. river gauges for load calculation) are included. For groundwater monitoring sites, the land use/cover surrounding the well can be given. Finally, the reference to the water body (river or groundwater body) completes the metadata.

Wastewater: MP concentrations in raw and treated wastewater including data from industrial and municipal treatment plants and settlements without any treatment. Further, data on sewage sludge can be found in this table set, as they share many metadata with concentrations in wastewater. Metadata associated with the wastewater samples are the sampling method (grab or composite samples), the sampling point in the wastewater treatment process and the flow volume at the time of sampling. The reference to the treatment plant was realized via a table for discharge points in accordance with the data structure of the EU Urban Waste Water Treatment Directive (UWWTD) reporting [27], which allows for one treatment plant to have several discharge points. The average discharge volume of the discharge points and the receiving water body are recorded and the discharge points are related to a treatment plant. The table on treatment plants further holds metadata regarding the catchment it is serving (industrial or municipal, number of inhabitants and share of combined sewer system), the design capacity (in population equivalent) and the implemented treatment steps and technologies.

Stormwater runoff: Concentration of MP in stormwater runoff either in combined sewer overflows or in stormwater outlets in separate sewer systems are collected together as a basis to quantify emissions from sewer systems. The metadata associated with the stormwater runoff samples includes the exact sampling point (before or after treatment facilities like retention ponds or soil filters), the sampling method, and the runoff volume during sampling. For the sampling sites, besides

the location information about the type of sewer system, the connected catchment (inhabitants, total area, share impervious, impervious and connected, industrial and traffic area), annual runoff volume, related precipitation gauge, mean annual precipitation and storage volume of the treatment facility.

Atmospheric deposition: MP contamination in atmospheric deposition was mostly reported as concentration in bulk deposition samples. However, one study used a different sampling approach, and therefore, the raw data were imported as deposition rates. The metadata for deposition samples includes the collection period, sampling method, precipitation amount included in the sample (either by measuring the sample volume and relate it to the sampler inlet diameter or by recordings from a nearby rain gauge) compared to the annual precipitation sum, and the solid content in the sample. The sampling site metadata include the location, the long-term mean annual precipitation sums and information about related (nearby) precipitation gauges.

Soil: MP concentrations in soil samples were available as single sample measurements as well as from analysis of composite samples combined from multiple sampling locations sharing some common properties. The metadata for soil samples includes information on the sampled soil layer (qualitatively as "humus cover" or "top soil" and the sampled depth section below the surface), sampling method, dry matter and organic carbon content of the sample, soil texture (qualitatively like"loamy sand" or by volume percent of each size fraction), and the soil horizon name (original value and master horizon according to [42]. The metadata for the sampling site, besides the location, contains information on the genetic soil type (original value and WRB reference soil group according to [42]) and the land use on the soil. All land use information was mapped using a controlled vocabulary taken from the CORINE land cover classification system [46].

The inventory was designed to include original monitoring data whenever possible, as well as data published in an aggregated form, such as in scientific publications and other technical reports. Therefore, for each environmental compartment different tables were adapted to include either single measurements or temporally or spatially aggregated values, such as statistically aggregated measurements from grab samples and measurements of time-integrated or space-integrated composite samples.

Additional tables contain metadata on the MP (names, identifiers) and the data sources (data owner, applying license) and are referenced from every dataset in the main tables (concentrations).

To ensure high data quality, several checks were built into the database as unique or check constraints to allow only the import of datasets with consistent metadata. This should be demonstrated with a few examples. Firstly, if a concentration measurement is marked as below LOD, a value greater than 0 and smaller than the LOQ value must be supplied for the LOD, and the measurement must also be marked as below LOQ. Secondly, sample or sampling site identifiers must be unique in the table to avoid having the same sample or site reported multiple times with different metadata. Finally, where the beginning and end of a sampling period must be supplied (e.g. for atmospheric deposition samples), the sampling end must be later than the beginning of sampling to avoid negative or 0 sampling durations.

Data collection

Data from all accessible sources in the DRB were collected, including surveys from national authorities, transnational monitoring programs (e.g. TNMN and JDS), as well as national and international research projects. The data collection focused on priority substances, priority hazardous substances, and other substances regulated under the EU directives 3 and 4, together with DRB-specific pollutants [61] and substances nominated as such, EU watchlist parameters [31], and substances selected for the monitoring programme conducted in the Danube Hazard m3c project [23]. In addition to the MP concentrations, the study also collected data on other water quality parameters that support data interpretation, such as suspended solids concentrations. Data mainly from the years 2008–2020 were requested and later updated, especially from the results of the Danube Hazard m3c project itself. Of the 14 countries with a significant portion of their territory within the DRB and contracting parties of the ICPDR, eight were represented in the project as project partners (AT, BG, HR, HU, ME, RO, SK, SI). However, only a few partner organizations were themselves the responsible institution with direct access to monitoring data (AT, ME, RO, SK). For three countries (HR, HU, SI), data were received via associated strategic partners, whose work was not directly funded by the project and data delivery was voluntary. For three further countries (BG, DE, RS), project partners successfully managed to receive data.

Data were mainly received in MS-Excel and comma separated value files and were read into R [57] with the packages openxlsx [58] and data.table [21], in order to harmonize and inspect them. The format and the vocabularies for the metadata were unified (e.g. using the same substance identifier for data from different data sources) and data and metadata checked for plausibility. Then data were imported into the database using the RPostgreSQL [17] or RPostgres package [65].

Data retrieval and evaluation

In the PostgreSQL database, data from different tables (e.g. for measurements, samples and monitoring sites) were combined in meaningful thematic database views. Database views allow to define a new virtual database table that consists of selected data from different existing tables joined together by common key columns. Such views were used to access the data for evaluation. While data for groundwater, river water and river SPM were collected in one set of tables, they are presented to the user in different views, to make them more easily accessible. The same applies for data on wastewater and sewage sludge, which are stored in the same tables, but presented in different views. Needed information for evaluation was selected and downloaded from the views by sending SQL queries from the R environment to the database and thus retrieve the results of the query for further evaluation directly in the R data analysis environment.

When dealing with MP concentrations, it is often the case that a significant portion of the values are censored, falling below the LOQ or even below the LOD. Therefore, it is essential to use appropriate methods to handle these cases during data evaluation to obtain the maximum information while avoiding introducing bias in the statistical analyses [37]. To calculate summary statistics for boxplots, the method "regression on order statistics" (ROS) was applied [38] using the NADA R package [48]. This method generally requires single measurements as input. To avoid losing all information included in composite samples and measurements that are only available as aggregated data, we adopted the following approach: we treated composite samples in the same way as grab samples, which may underestimate their representativeness but at least partly utilizes the information they contain. For aggregated data from statistical aggregation of single values, we considered their mean value, by treating calculated mean values below LOQ the same way as we treated censored measurements from grab samples.

Statistical testing of relationships between metadata and concentration levels is necessary to investigate which metadata can be used to explain different MP concentration levels in an environmental compartment. These metadata can subsequently be used for better inter- and extrapolation of the available monitoring data into unsampled temporal or spatial regions for emission modelling. To test the correlation of a numerical variable with the partly censored concentrations, Kendall's τ was used, whereas a generalized Wilcoxon test also known as Peto–Peto test [38] was applied to test the influence of a categorical variable on the partly censored concentrations (cenken/cendiff function from the NADA R package). To be able to investigate also multivariate relationships with multiple independent variables and multiple substances concentrations as dependent variables, we searched for an ANOVA-like tool that can account for censored data (significant share of data below LOQ) and is not based on assumptions like multivariate normality, which are often not met by the concentration data. Following Helsel 38, ranks of u-scores were calculated substance-wise from the concentrations to handle censored data using the NADA2 R-package [43]. Then a dissimilarity matrix using Euclidean distance was derived and fed together with the independent variables into a permutational multivariate analysis of variance (PERMANOVA, [5] and into PERMDISP [6]), a multivariate extension of Levene's test. Both procedures were conducted using the vegan R package [54]. PERMANOVA does not require input data to meet any given distribution and allows classical partitioning with tests and estimation of sizes of main effects, interaction terms, hierarchical structures and mixed models but it is sensitive in unbalanced sampling designs to inhomogeneity of dispersion [7]. Therefore, the homogeneity of dispersion was investigated with the PERMDISP-test before applying PERMANOVA. To cope with unbalanced sampling designs, Anderson et al. [9] developed an adapted test metric, which is less influenced by inhomogeneity of dispersion. Unfortunately, these algorithms are not yet implemented in the R programming environment and therefore could not be applied in this work.

To examine through substance-wise post hoc test how the median values differ in the groups defined by the metadata, a one factor permutation test (cenpermanova from NADA2) was applied. Correlation between the independent variables was checked by ANOVA for mixed types (numerical and nominal) and with Cramer's V [18] for nominal variables using R package rcompanion [51].

Results and discussion

To demonstrate the capabilities and usefulness of such a database of MP concentrations, this section presents and discusses exemplarily selected results regarding the assessment of data availability, identification of important emission pathways, and origination of model input data for emission models. Subsequently, this section is completed with the main insights and lessons learned regarding the data management process and the structure of the database, in view of its further development as fully operative tool.

Data availability

In total 10.7 M concentration measurements in over 383 k samples from about 25 k sampling sites are included in the database. Figure 1 illustrates the distribution of the measurements across the number of substances and the environmental compartments investigated. It is evident

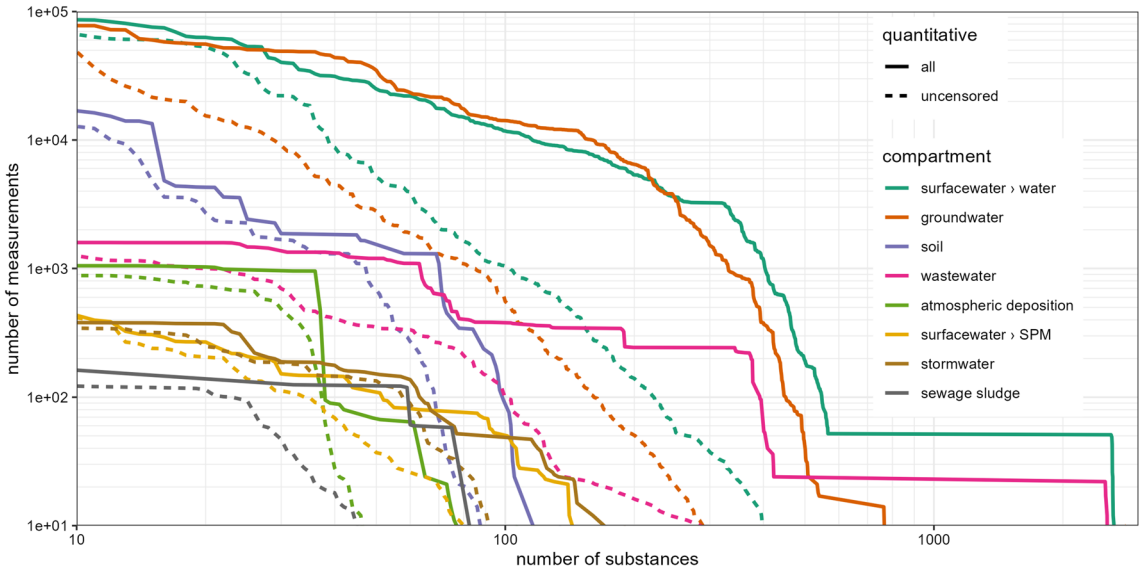

Fig. 1 Content of the concentration database for the DRB: Number of concentration measurements versus number of substances for different environmental matrices. Full lines represent all measurements, dashed line only those above LOQ

that a high number of measurements are available for only a small number of substances (30–40), while only very few samples were analysed for a very high number of substances (> 2000) in surface and wastewater. This can be easily explained by the different monitoring approaches implemented in the DRB. One the one hand, in accordance with the requirements of the WFD and its daughter directives, approximately 40 parameters must to be monitored. For these parameters, a significant amount of data for ground and surface water stems from surveys conducted by national authorities. On the other hand, within the JDSs, organized by the ICPDR and supported by several research laboratories, a few selected locations along the Danube and since JDS4 also selected wastewater treatment plants (WWTP) are sampled once every 6 years. This results in a limited number of samples being analysed for a large number of chemicals, which increases over time as analytical capabilities develop.

The matrix with the most available results is surface water, followed by groundwater. For soil, noteworthy numbers of measurements are available for a few substances. The spectrum of analysed substances is wide for wastewater, but the total number of measurements is rather small (less than 2000). All other matrices (stormwater runoff, atmospheric deposition, sewage sludge and SPM) have far fewer measurements and measured parameters.

The data availability over time (Fig. 2) indicates that monitoring of surface water and groundwater is a relatively continuous process with slightly increasing number of measurements and parameters being investigated. Untreated wastewater numbers have started lower but are strongly increasing, while data for treated wastewater are only available for some years. This is mainly due to concentration data derived from reported loads being excluded from the evaluation (please refer to the last section of "Results and Discussion" chapter for further explanation). For soil there is continuous monitoring, but no clear trend over time was identified. Monitoring of atmospheric deposition and stormwater runoff is sporadic and mostly limited to dedicated research projects, with no discernible trends identified. The numbers for all compartments drop in 2020, which may be due to the fact monitoring data were not yet available when data collection for the database started in March 2021 or could be an effect of the COVID pandemic.

The data availability in the DRB varies greatly between different countries (Fig. 3). Germany and Hungary have a high number of measurements from various environmental compartments, while for all other countries data from fewer environmental compartments are available and the number of measurements is usually low, except for surface water. The number of parameters has a high variation, with countries and compartments investigated in the JDS4 including very high number of substances, while other data sources contain much fewer substances. Bosnia and Herzegovina (BA), Czech Republic (CZ) and Moldova (MD) were not actively involved in the Danube Hazard m3c project and therefore national monitoring data are not yet included, except those available from transnational monitoring activities (JDS4 and TNMN).

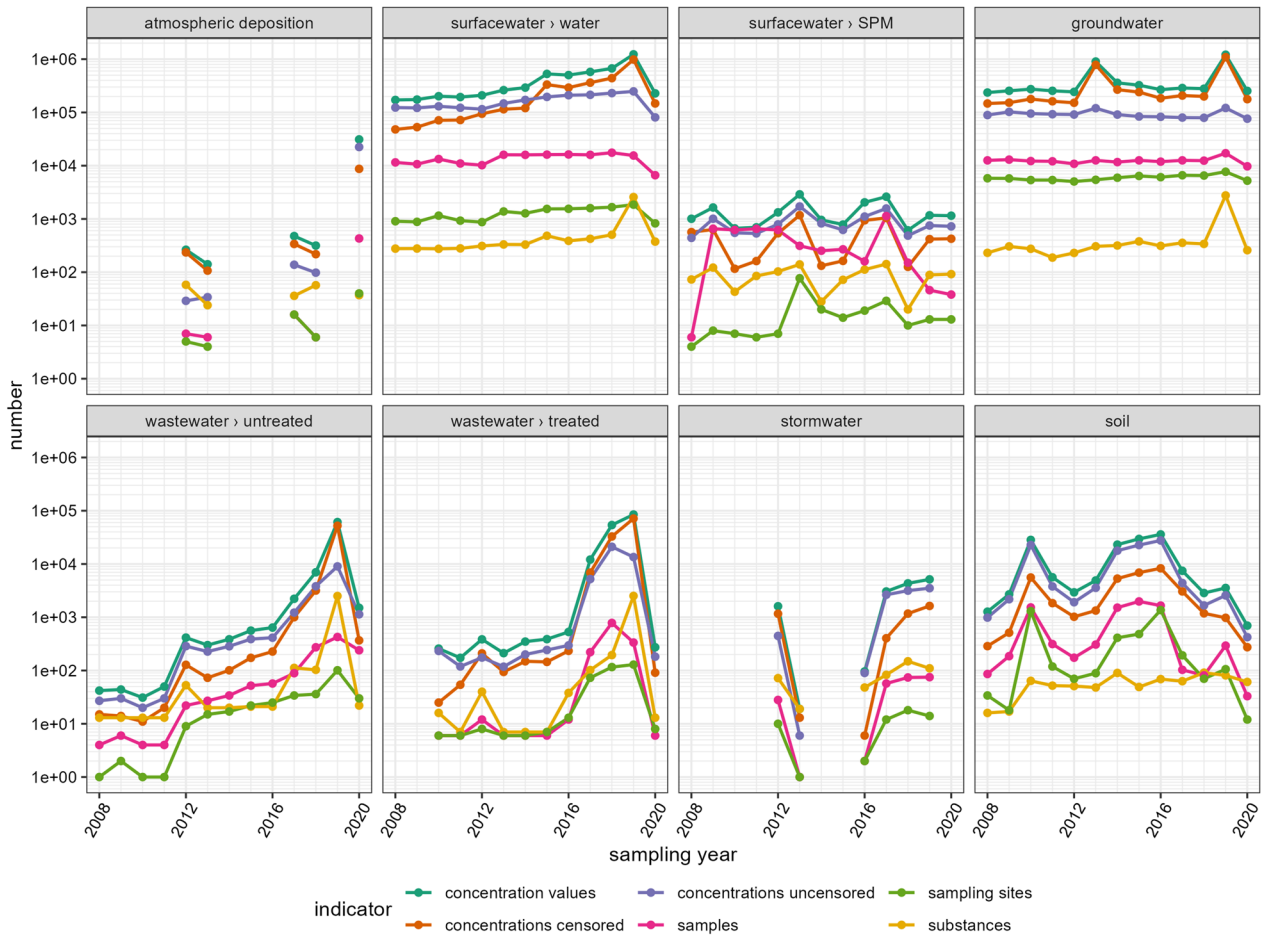

Fig. 2 Availability of monitoring data for the years 2008–2020 for different environmental compartments in the database for the DRB

The spatial coverage of the data is generally very heterogeneous (Fig. 4). Even for surface water and heavy metals, where the highest numbers of monitoring data are available, the spatial coverage is incomplete, with gaps mainly in the southeast of the basin (Fig. 4:A). Mercury concentrations, which are difficult to quantify in the low ranges occurring in surface waters, are not measured quantitatively in many places, so that the data density is even lower, when only looking at areas with quantitative results (Fig. 4:B). Generally, the data availability is higher along the Danube main river compared to the tributaries. This holds also true for carbamazepine, an anticonvulsant pharmaceutical compound that is hardly degraded by conventional wastewater treatment [15] and for azoxystrobin, a fungicide used in agriculture and listed on the new surface water watch list under the WFD [36] (Fig. 4:C-D).

In terms of spatial coverage, it can be concluded that even for surface waters, which are the best-investigated compartment, and for well-known substances like mercury, data coverage is very scattered. There is only a certain, perhaps sufficient density of monitoring sites along the Danube. Targeted monitoring approaches such as the JDS can help to close data gaps. However, if aiming to understand the emissions and fate of substances in the entire basin, the tributaries must also be covered.

Identification of main pathways

The proportionate contribution of specific pathways to the overall emissions of MP into surface waters depends on the volume of the transport medium (discharge, sediment input or precipitation) and the level of contamination within that medium. Analysing either of these factors by comparing concentration levels across different environmental compartments provides initial insights into the predominant routes through which substances are emitted. Additionally, contrasting the pollution profiles of different types of MP can help identify specific indicator substances. If pollution levels of these indicators are elevated in one compartment, it signifies pollution primarily via the related pathway. Figure 5 compares the concentration levels of six selected MP from different

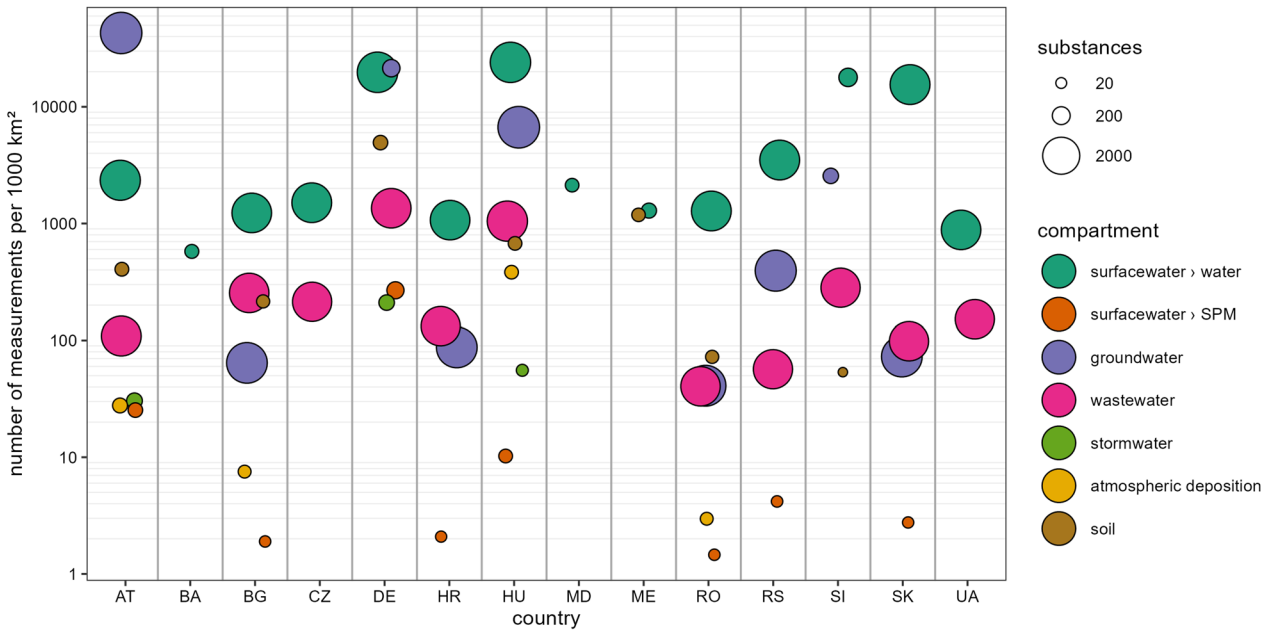

Fig. 3 Data availability on country level. All countries with > 2000 km^2 area in the DRB (contracting parties of the ICPDR) are shown. The number of measurements is normalized with the area of the country in the DRB. Where no data for an environmental compartment are available, no point is shown

substance groups in bulk atmospheric deposition (atm. deposition), groundwater, river water, stormwater runoff (from storm sewers in separated sewer systems and from combined sewer overflows in combined sewer systems), effluent from municipal WWTP, topsoil, SPM in rivers and sewage sludge. Although concentrations vary in many cases over several orders of magnitude, the median concentrations provide a rather clear indication of where contamination levels are higher and where less contamination is found:

Benzo[a]pyren (B[a]P) is a polycyclic aromatic hydrocarbon (PAH), well known as carcinogenic air and water pollutant mainly produced during incomplete combustion of organic substances. Therefore, residential heating with wood or coal fire and industrial emissions are significant sources [59]. Emissions from these sources are mainly distributed via the atmosphere and deposited onto surfaces. In water, B[a]P has a strong tendency to adsorb to particles. Figure 5 shows that highest B[a]P concentrations are detected in stormwater runoff and still significant concentrations are found in atmospheric deposition, while comparable low concentrations are measured in treated wastewater. The latter can be explained by the efficient removal of B[a]P through adsorption to sewage sludge. Concentrations in surface and groundwater were mostly below LOQ and are therefore not shown as boxplot. When comparing concentrations in solids, levels in sewage sludge and surface water SPM are similar, with a higher variability in river water SPM, while concentrations in topsoil are one order of magnitude lower. Therefore, it can be suspected that stormwater runoff is the main emission pathway for B[a]P in urban areas, while soil erosion might gain importance in rural areas impacted by agricultural erosion. Emissions via WWTP effluents might be negligible. The importance of stormwater outlets is supported by results from emission modelling [13, 33].

For mercury, Fig. 5 shows total concentration levels in the same order of magnitude in atmospheric deposition, river water, stormwater and treated wastewater, with highest total concentration in stormwater runoff, followed by atmospheric deposition. Concentrations in groundwater were not quantitatively measured above the LOQ with the applied methods and are therefore probably lower. As for B[a]P, the direct deposition onto water surfaces and wash-off from impervious surfaces are important pathways for mercury river pollution. However, in comparison to B[a]P, concentration levels in deposition are significantly higher and thus the direct deposition on water surfaces gains importance. It is thus less clear if there is a dominant pathway. This is also supported by the content in different solids: concentration levels in soil, river SPM and sewage sludge are around the same level, with slightly higher concentrations in sewage sludge. These findings are consistent with the literature, which identifies air pollution from coal combustion and

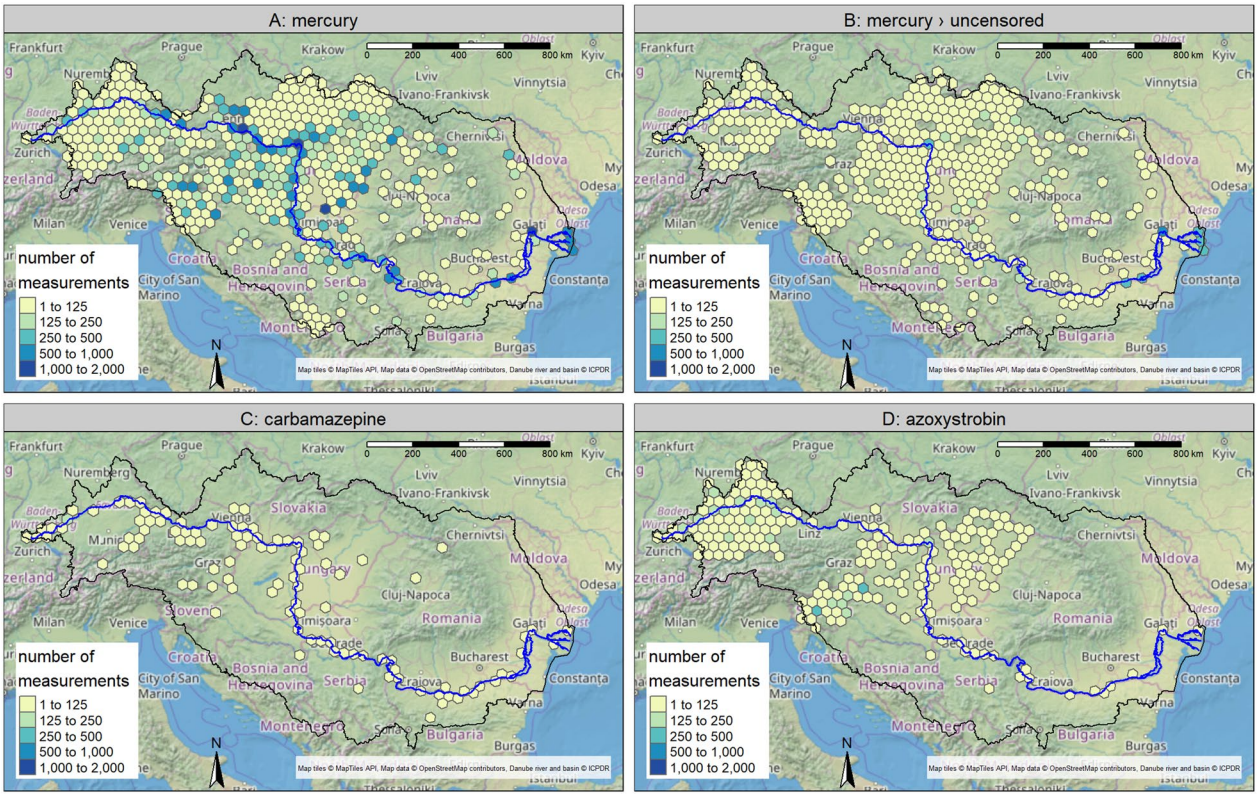

Fig. 4 Spatial availability of measurements in surface water (water samples) for selected substances: mercury (**A, B**), carbamazepine (**C**), azoxystrobin **(D**). Number of measurements are summarized on a regular raster, in areas without visible raster cells no measurements are available. To highlight the relevance of sensitive analytical methods for mercury, additionally to the total number of measurements in panel **A** the number of measurements above LOQ is presented in panel **B**. Where concentrations in total and dissolved matrix were available, both were counted

cement production as the primary emitters of mercury into the atmosphere [26]. Through air transport and deposition, mercury is contaminating sealed areas, topsoil and water bodies and thus several different pathways might be relevant [32, 55].

Copper (Cu) is an element used in many technical processes but also occurring in nature. High emissions can be expected from traffic due to the use of Cu in brake-systems, resulting in significant brake-wear emissions. In Fig. 5, the highest total Cu concentrations can be seen in stormwater runoff, followed by treated wastewater, atmospheric deposition and surface waters. This is in line with the above-mentioned emissions from traffic. As the database is very broad for Cu, one can also see some very high concentrations in river and groundwater, probably related to areas with geogenic elevated concentrations and associated mining activities. Median total Cu concentrations are lowest in the groundwater compartment. With respect to solids, the highest Cu concentrations are found in sewage sludge, followed by river-SPM and soil, with significantly lower concentrations. Soil concentrations show a high variability, where the high outlier may again be attributed to areas with elevated geogenic background concentrations and mining activities. These findings indicate that for Cu stormwater runoff is a major pathway in urban areas and that areas with elevated geogenic background concentrations and mining activities need to be explicitly included in emission models to obtain meaningful results.

Diclofenac is a widely used anti-inflammatory drug. Thus, emissions can be expected mainly via sewage. Diclofenac is not easily degraded during conventional wastewater treatment [50]. Figure 5 shows highest Diclofenac concentrations in treated wastewater, followed by concentrations in stormwater (primarily due to data from combined sewer overflows), which are one order of magnitude lower, and river water, which has again significantly lower concentration levels. The lowest concentrations were detected in atmospheric deposition and groundwater, with levels 4 orders of magnitude lower than in treated wastewater. Sewage-related pathways, namely WWTP effluent and combined sewer overflows, are the dominating pathways of diclofenac emissions into surface waters.

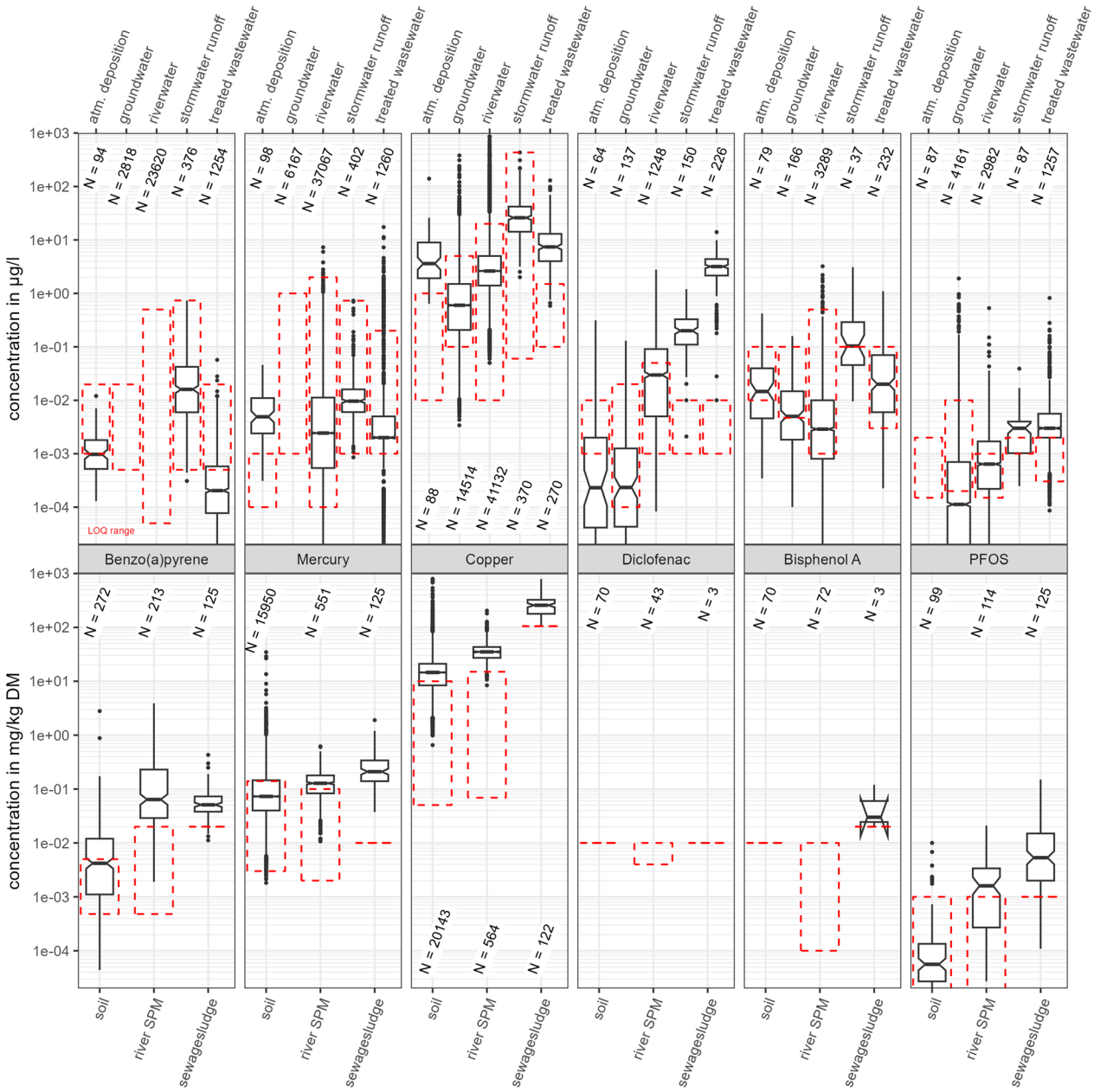

Fig. 5 Concentrations of selected MPs in different environmental compartments. Concentrations are shown as box-whisker plots with values below LOQ imputed by means of ROS under assumption of a lognormal distribution. If more than 80% of observations were below LOQ, no boxplot is shown. Notches in the boxes indicate roughly 95% confidence interval for comparing medians. The range of the LOQ is indicated by red dashed box. The number of observations (N) is shown as annotation above or below the boxes. The numbers underlaying this figure can be found in Additional file 1

Bisphenol A (BpA) is a chemical used as a monomer for the production of polycarbonate and epoxide resins, from which remaining monomers can leach during the use phase. BpA is also used in its original form, e.g. as an antioxidant in break fluids [47]. Its endocrine-disrupting impact [44] makes its occurrence in the aquatic environment a matter of high concern. Here, highest concentrations of BpA are found in stormwater runoff, followed by treated wastewater and atmospheric deposition (Fig. 5). Lower concentration levels are found in groundwater and river water, with lower median concentrations in river water than in groundwater (although the data base for groundwater is still rather small). In solids, only in sewage sludge some values above LOQ can be

seen. With a solubility of 300 mg/L in water [60], sorption to solids is not a decisive process for fate and transport of BpA. Thus, the main pathways for BpA is probably stormwater runoff or treatment plants effluents, depending on the sewer system and emitted water volumes.

Perfluorooctanesulfonic acid (PFOS) is a very persistent substance that was used in many applications until it was regulated under the Stockholm convention. PFOS remains present in the environment as legacy pollution, as well as in consumer products and firefighting foams found in the technosphere, which may still emit PFOS. Furthermore, PFOS can derive from degradation of precursor substances, especially during conventional wastewater treatment [52]. Figure 5 shows that the median concentration levels in treated wastewater and stormwater runoff are similar. In contrast, surface waters have significantly lower median concentrations, and in groundwater they are nearly one order of magnitude lower. In atmospheric deposition, the concentrations were not detected above the LOQ in most cases. Groundwater concentration shows the highest number of observations and also a very high variability with extremely high values, which are even higher than in treated wastewater. This can be attributed to local groundwater contamination, which in many cases is legacy pollution caused by former firefighting foam applications, e.g. in training fields at airports [14]. When looking at the solids, highest contamination is found in sewage sludge, followed by river SPM. Concentrations in soils are rather low and, as a result, not a significant pathway of river pollution. When modelling emissions, it is important to consider that in most catchments wastewater and stormwater runoff are the primary pathways. However, in some areas exfiltration from contaminated groundwater may be more significant and should also be included in the models, as described, e.g. in Kittlaus et al. 45.

Deriving input data for emission modelling

Emission models for regionalized pathway analysis, such as the MoRE model [34], require concentrations as input data for different pathways. Ideally, these data are temporally and spatially distributed to best represent the differences in pollution occurring in time and space. But as data on concentration of MPs in different environmental compartments are scarce, often one constant concentration value is applied for a whole river basin and over multiple years. Thus, every statistically significant differentiation between concentrations in space or time or by attributes of the emission pathway (e.g. applied treatment technology in WWTPs or land use stratification of top soils) has the potential to improve the results of the model application.

A first example of how this inventory of concentrations can support emission modelling from municipal WWTPs

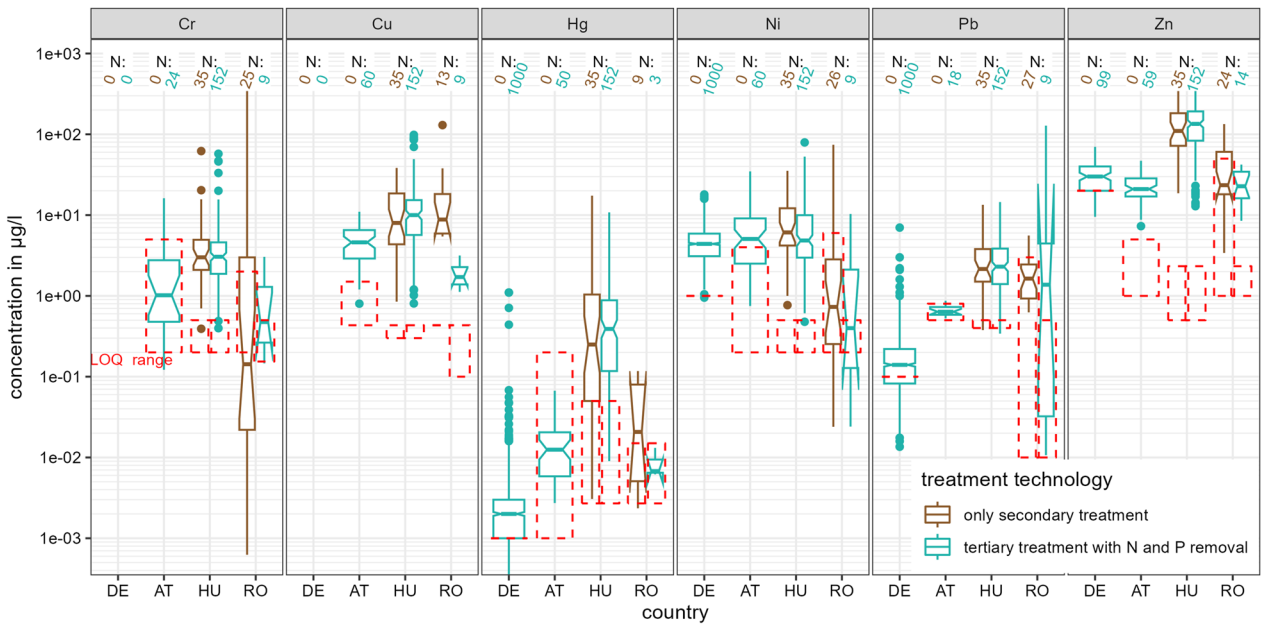

Fig. 6 Heavy metal concentrations (total) in treated wastewater from municipal WWTP in different Danube countries with different treatment technologies implemented. Concentrations are shown as box-whisker plots with values below LOQ imputed by means of ROS under assumption of a lognormal distribution. Notches in the boxes indicate roughly 95% confidence interval for comparing medians. If more than 80% of observations were below LOQ, no boxplot is shown. The range of LOQ is indicated by a red dashed box. The number of observations (N) is shown above each boxplot. The numbers underlaying this figure can be found in Additional file 1

is through the examination of heavy-metals concentrations in WWTP effluent across different countries (Fig. 6). Quite large differences in effluent concentrations of municipal WWTPs can be observed for certain metals. For instance, the median concentration of total Hg and Zn in Hungary is approximately one order of magnitude higher than in Germany, Austria, and Romania. To examine whether this discrepancy is due to varying levels of treatment technology, the plot also indicates the type of treatment used. It is evident that the treatment technology applied is not the determining factor. It is worth noting that the median concentrations of total Hg and Pb in Germany are one order of magnitude lower than in other countries, while this is not the case for Ni and Zn. The data were thoroughly checked for quality problems, but no issues were found that would affect comparability. However, the reason for these deviations could not be identified with the available data, and further research is needed.

These significant differences in concentrations at national level can be seen as indicator that, when preparing input data for emission modelling on a larger scale, it is important to use different input data for municipal WWTP effluent in different countries. Furthermore, the low number of concentration values in all countries except for Germany is a strong indicator of the need for further monitoring of WWTP effluents with suitable methods to broaden the data basis.

As a second example, this study investigated the concentrations of PFAS in the effluent of municipal WWTPs. A multivariate data analysis was conducted to determine the factors that influence the PFAS concentrations in the effluent. The dataset used in the analysis contained concentrations of PFPeA, PFHxA, PHpA, PFOA, PFNA, PFDA, PFBS, PFHxS and PFOS in 1036 effluent samples from three data sources [23, 49, 53]. The following independent variables were examined: size of the treatment plant given by 5 classes for the design capacity (in population equivalents (PE), see Fig. 7), treatment steps applied (only "secondary treatment" or "tertiary treatment with N and P removal"), country of the WWTP (AT, BG, CZ, DE, HR, HU, RO, RS, SI, SK) and time of sampling, which was processed into sampling year (2017–2022) and season (summer, winter). Unfortunately, the availability of data in the different groups resulting from the above-mentioned independent variables is very unbalanced, with many combinations of independent variables without any samples and some combinations with high number of samples. Due to this unbalanced sampling design, a homogeneous dispersion of the values in the groups defined by the levels of the independent variables is a precondition to test for differences in location of the groups [8, 9]. With the PERMDISP test (9999 permutations), the data were tested for homogeneity of dispersion and this could be rejected with high significance ($p < 0.001$) for all variables. Only after pooling the classes of the

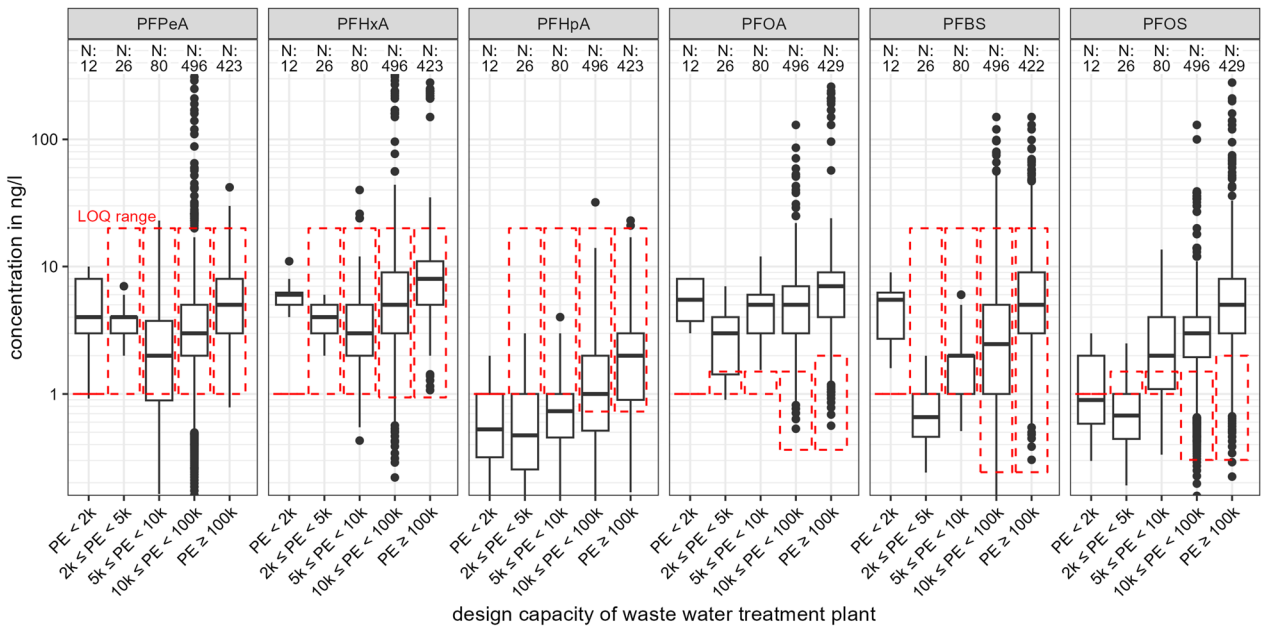

Fig. 7 PFAS concentration in effluent of municipal WWTP in the DRB depending on the size of the WWTP given in PE. Values below LOQ (range given as red dashed boxes) were imputed using ROS under assumption of a lognormal distribution. The number of observations (N) is shown above each boxplot. The numbers underlaying this figure can be found in Additional file 1

WWTP design capacity into capacity below 100 kPE and above 100 kPE, similarity of dispersion of these two groups could not be rejected ($p = 0.67$). Following this, it was tested if the mean effluent concentrations between WWTPs above and below 100 kPE are significantly different and this was confirmed with PERMANOVA (9999 permutations, $p < 0.001$). A substance-wise post hoc one-factor permutation test (cenpermanova from NADA2) indicated that for all substances that showed significant ($p < 0.05$) differences in the mean concentration (PFHpA, PFOA, PFBS, PFHxS, PFOS), the effluent concentrations from WWTPs with a capacity above 100k PE were higher than those from the smaller treatment plants (the test results and group-wise mean concentrations are listed in Additional file 1).

The significant different concentration levels in effluent of WWTPs above and below 100 kPE can already be used as input for emission modelling. Figure 7 suggests the potential existence of further patterns that are not statistically significant within the current unbalanced data basis, but which might be interesting to further investigate after broadening the data availability. Specifically, the concentration levels in small plants seem to be higher than in medium sized plants. For those size classes for which a high number of measurements is available, the concentrations spread over a wide range, in some cases 3 orders of magnitude, meaning that the groups are very inhomogeneous. This might indicate that the main emissions are not caused by sewage from households and widespread commercial activities, but rather by specific commercial or industrial activities that exist in some of the WWTP catchments and not in others. To improve the spatial precision of the emission models, it would be necessary to identify these critical activities and include them in the modelling.

As a third example, the concentration of selected MPs in top soil, which serves as input data for the emission pathway "soil erosion", should be tested in terms of which metadata might be suitable predictors for extrapolation to unsampled areas. MPs from different substance groups, which are expected to be relevant soil pollutants and where an adequate amount of measurements were available in the database, were selected: mercury as an example for heavy metals, a low molecular weight PAH (phenanthrene) and a medium molecular weight PAH (fluoranthene), hexachlorobenzene (used as pesticide in the past but that can also derive from breakdown of other chlorinated organic substances), and a long-chain PFAS (perfluorononanoic acid, PFNA) with a comparable high toxicity, indicated by a high relative potency factor ($RPF = 10$) in the proposed new EU environmental quality standard for water [16] final based on [12]. Regarding the available metadata, information about land use on the soil, time of sampling, country where the sampling took place and the data source were available. Some datasets further contain information on the soil texture and genetic soil type, grain size distribution and organic carbon content, but these data were too scattered for a statistical investigation. Nevertheless, also those variables which were available for all samples (land use, sampling year, country and data source) could not be tested with multivariate procedures, as none of the selected substances are included in all data sources (this is only the case for some commonly measured heavy metals) and thus no complete cases remained for the statistical analysis. Therefore, only univariate procedures could be applied and interaction effects could not be considered. Highly significant ($p < 0.01$) differences in concentrations of all selected substances between different types of land use on the soil were found using the Peto–Peto test (the detailed test results are reported in Additional file 1). This means that at least one land use shows statistically different soil MP concentrations than other land uses (Fig. 8). Figure 8 shows, however, that the available information on land use is not totally consistent. The general class "agricultural land" includes sampling sites with agricultural land use that could not be attributed to the specific sub-classes "arable land" and "pastures", such as for example "land principally occupied by agriculture, with significant areas of natural vegetation" or "heterogeneous agricultural areas with complex cultivation patterns". However, an even larger portion of data was available for agricultural areas, for which no detailed land use information was available. Therefore, classification needed to remain on this general level. For mercury, both investigated PAHs and PFNA, the top soil concentration in forests is higher than on arable land and pastures, while for hexachlorobenzene concentrations on arable land are significantly higher than in forest soils. This can be explained by the different input of the substances into the soil, mainly via ubiquitous atmospheric deposition in the first case and via pesticide application on arable land for hexachlorobenzene. In forests, substances deposited by the atmosphere are subject to a combing effect due to the enlarged surface area of branches and leaves. Additionally, these substances remain in the upper soil layer for longer periods than on arable land, where the soil is regularly mixed by the tillage. The different sampling depth for top soil sampling in forests (usually about 10 cm) and on arable land (usually about 30 cm) further explain the lower MP concentrations found for contaminants transported via atmospheric deposition in arable land compared to forest soils. Not easily explainable are the causes for the high hexachlorobenzene contamination in soil of seminatural areas. Further metadata variables (organic soil content, soil texture, year of sampling)

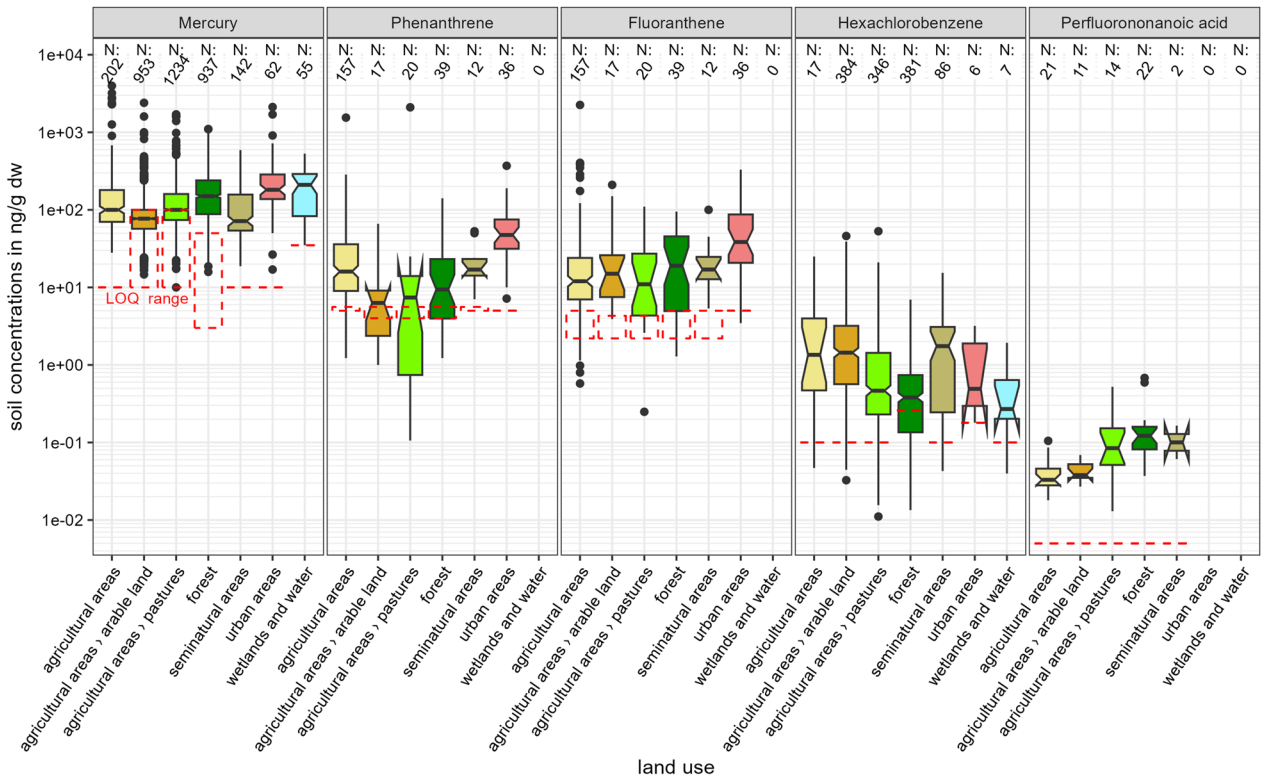

Fig. 8 Top soil content of mercury and selected organic MPs versus land use on the soil. Concentrations are shown as box-whisker plots with values below LOQ imputed by means of ROS under assumption of a lognormal distribution. Notches in the boxes indicate roughly 95% confidence interval for comparing medians. If more than 80% of observations were below LOQ, no boxplot is shown. The range of the LOQ is indicated by red dashed box or line. The number of observations (N) is shown above each boxplot. The numbers underlaying this figure can be found in Additional file 1

would be needed to shed light on the underlying drivers and to derive an improved regionalization model for soil concentrations not only depending on land use.

Insights regarding data management and database structure development

During the steps of data collection, data cleaning, and database setup, several insights and lessons were learned. These will be briefly discussed in the following section.

Data accessibility was a major issue during data collection, particularly when the data holding institution was not a project partner responsible for the data collection or when a different authority or department within the same authority held the data, as they did not originate from the water administration, such as soil concentrations. One reason for authorities holding back data was of financial nature, as they sell monitoring data as a strategy to cover their expenses for data management and processing. In other instances, they were hesitant to release the entire datasets as they typically only provide selected portions of data to interested users, rather than complete set. As a result, some authorities decided not to provide raw data, but only aggregated data, e.g. mean annual river concentrations instead of single measurements. These data were included in the database, but aggregation causes significant information loss in terms of gained understanding and capability of extrapolation. For instance, it is not possible to evaluate mean annual river concentrations with regard to differences between low-flow and high-flow situations. This hinders, among others, the reliable calculation of riverine loads, which are required for the generation of emission inventories.

In many cases, data were made exclusively available for use within the project. The providing institutions did not agree to include their data in the published database, not even if the origin of the data was properly documented. One reason given for not agreeing to republish whole raw datasets was the loss of the opportunity for further data corrections. This is however a controversial issue: as the data are not or only constrainedly distributed, errors in the dataset may not be discovered by potential users. One solution to this contradiction might be the setup of interfaces between different data bases, which allow for easy data updates whenever necessary. Even when data

owners agreed to publication, the process of reaching a data publication agreement or assigning a license for reuse was time-consuming. As an example of best-practice, we identified the Bavarian Environment Agency, which offers its data for download [11] and clearly licenses them under the Creative Commons Attribution 4.0 International License [19]. This facilitates the process greatly.

As already mentioned, emissions from WWTPs are reported as loads under the EU UWWTD. Therefore, in many countries they are only available in the form of reported loads. While water amounts were available to recalculate mean concentrations from these loads, the approach to include concentration data below LOQ as 0 during load calculation leads this reverse approach ad absurdum, because the higher the LOQ of the applied analytical method is, the more measurements fall below the LOQ and are included as 0 and thus the lower the back-calculated concentrations are. Therefore, even if these data were available, they could not be used for further investigations in this study and were excluded from the evaluation. In conclusion, we recommend to store and share data in a non-aggregated form.

During the collection and checking of the available data, it was observed that several responsible institutions lack the necessary tools and skills to handle large datasets, despite their high motivation. Many institutions still store and manage data in spreadsheets without implementing appropriate automatic quality checks. This increases the likelihood of human errors, such as incorrectly assigned units of measure for concentrations or omitted information about analytical limits and methods. If data are not stored in well-organized databases with controlled vocabulary, the available data may contain different terms to describe the same content. This could be due to different people using different words to describe a method or simply because of typos that were not noticed during data creation. Mapping heterogeneous content onto a controlled vocabulary is a laborious task that often requires a high level of expertise to determine which terms describe the same thing and where actual differences need to be documented. When data are derived from a well-designed database system, data mapping requires considerably less effort.

Concerning the database development process, it was observed that the database design and controlled vocabulary setup were discussed by a small team but mainly implemented by one project team member, while data import was shared among a few team members. Although this was the only way to quickly establish a functional system and fill it with data, it occasionally resulted in misunderstandings between the developer and the data manager. This led to a need for further data harmonization and adaptations of the naming and documentation in the database. When developing such a tool, which will be used by users with different backgrounds, it is essential to allocate time and resources for common discussion of the terminology and ideas, as well as for proper documentation and testing of the tool, to achieve consistent results.

With the ongoing import of further data, a better understanding of available data and metadata was developed and the data structure and controlled vocabulary were adapted and extended. For example, initially for all environmental compartments, separate tables for single measurements and aggregated measurements were created. But for some compartments, such as atmospheric deposition and stormwater, no aggregated data were received, making these tables superfluous. On the other hand, controlled vocabularies, which were not imported from elsewhere, grew during data import as some data could not be mapped on the available values and were thus added to the vocabulary. From time to time, a reorganization of the controlled vocabulary was necessary, including updating the already imported data. To make sure that during this process of evolving classification the original value is not evolving into something incorrect, it is recommendable to additionally keep the value as contained in the raw data for later check-up.

The probably most difficult but also most interesting topic regarding the database design is the identification of useful metadata and the selection or development of controlled vocabulary for these metadata. The most important metadata for concentration measurements is the substance analysed. Already here, several challenges are faced when combining data from different data sources: MP often are reported using different names, may it be a product brand name, the standardized chemical IUPAC name or another possibly shorter common name. Fortunately, nowadays in most cases a CAS number is reported along with the substance name, which helps identifying the chemical substance more easily. However, in some studies outdated CAS numbers are found or the CAS number is given in a report, but not directly together with the tabulated data. For some substances, no CAS number is available, as they are only known from environmental screening methods and the substance itself is not characterized yet (e.g. some pesticide metabolites found in the JDS). In these cases, the unique identifiers from the NORMAN network database were used. On the other hand, for some MP several forms of the compounds are reported, which are chemically different, either because differences in the structure or some carbon chains occur or because different salts of an acid or base are used as standard for the analysis, but which—from the perspective of water pollution—somehow describe the same

pollution or are actually not differentiable as they occur dissociated in the water matrix. The reported CAS number is anyhow different in such cases, depending on the exact standard used for calibration in the laboratory. Here, a high expertise in chemistry (or even toxicology) is necessary to decide which substances can be evaluated together as one MP and where differentiation is necessary. A reliable database on substance identifiers, names and relations is of great assistance for such questions, and a database as presented here can help to set up a knowledge base regarding such issues.

Some examples should showcase problems regarding substance identification and solutions implemented in the database. Nonylphenol is a substance used in industry, but also deriving from nonylphenol ethoxylate degradation in the environment, and it was identified as priority hazardous substance by the WFD. Actually, it is a group of substances with the same chemical formula $C_{15}H_{24}O$ but different structures, and in industry it is mostly produced as a mixture of different isomers and structures. Chemical laboratories apply different standards for instrument calibration and therefore reported data include nonylphenol (CAS 25154-52-3), 4-nonylphenol (CAS 104-40-5) and branched 4-nonylphenol (CAS 84852-15-3). These were separately collected in the database, but the question arises as to whether or not these substances should be grouped together or not for evaluation purposes. Another interesting example is Mecoprop, an herbicide used on green roofs in the sealing membrane to protect it against penetration from roots. The product Mecoprop is applied as a 1:1 mixture (CAS 93-65-2) of two isomers, one of which is herbicidal active (CAS 25333-13-5) and one not (CAS 16484-77-8). Unfortunately, in the received data a fourth CAS number occurs (CAS 7085-19-0), for which the exact relation to the others is not clear to the authors, even if a substance info page of the European Chemicals Agency exists [24].

A pre-existing controlled vocabulary was applied in the case of land use metadata (for soil and groundwater sampling sites), taken from the CORINE land cover dataset and well documented in Kosztra et al. 46. This has two advantages, namely it can be assumed that all land uses occurring in Europe can be mapped to this classification system and missing information could be generated by analysis of the geodata based on the sampling site locations. Nevertheless, this classification system is a combination of land use and land cover information and better data models are under development [10].

Conclusions and outlook

By combining concentration data from different data sources in a common database for the DRB, several new insights could be gained.

Data availability is very heterogeneous for different substances and environmental compartments. To achieve a more balanced data availability, it would be necessary to harmonize monitoring networks and investigated substances. This would allow to investigate the spatial distribution of pollution. While surface and groundwater are rather well covered by existing data, many countries lack data for other environmental and technical compartments. This is either due a lack of monitoring or inaccessibility of data. We propose to develop a harmonized monitoring program for the DRB not only covering river monitoring along the Danube river itself and a few large tributaries (as currently implemented by the TNMN and the JDS), but also river monitoring in many tributaries and in other environmental and technical compartments. For such a harmonized monitoring approach, it would be important to collect metadata of potential sampling points in advance and ensure a well-balanced sampling design that facilitates statistical evaluation.

During the collection of available data in the basin, deficits were identified in the data management systems and data literacy within the data holding institutions. To address these issues, it is strongly recommended to use well-designed database systems to manage monitoring data. The importance of proper handling of data cannot be overstated, especially when considering the high costs invested in sampling and chemical analysis. It is crucial to have a good understanding of technical aspects and licenses and ensure that the data are easily accessible for reuse. Capacity building regarding data literacy in, e.g. national water administrations and lobbying for making environmental monitoring data FAIR [66] can help to raise the benefits of the monitoring data for society.

Applying sound methods for handling censored data is especially important when working with MP concentrations. To be able to apply these methods, it is of crucial importance that the single concentration values together with the analytical thresholds (LOQ) are available, as these methods do not work on previously aggregated data.

When combining data from different data sources, metadata required for extrapolation is often missing, preventing the reuse of data for this purpose. Where metadata exists, the used vocabularies may differ, substances may be named differently and different CAS numbers (some of which may be outdated) are used as identifiers for the same substance. Mapping such vocabularies from source datasets onto a target vocabulary is a time-consuming task requiring qualified personnel. Nevertheless, this step is necessary to be able to evaluate data collectively. A potential solution would be to provide harmonized controlled vocabularies on an EU level (e.g. the EIONET WISE vocabularies, [28] and to implement

them directly in national and transnational monitoring programmes. Application of artificial intelligence could also aid in expediting these tasks in the future.

Data necessary for quantifying emissions to surface water do not solely originate from the water discipline. Other data about soil pollution and atmospheric deposition are also required. This data is usually collected by different departments or organizations and the metadata collected alongside the data might not be sufficient for answering the questions related to water pollution. Therefore, a database system with a well-chosen selection of useful metadata can aid in communicating metadata needs and improving transdisciplinary cooperation and data valorization.

The current version of the database is a proof of concept and can be used by experts for scientific purpose. It holds the potential for many further investigations, such as expanding the substances analysed or conducting data analysis on other pathways, such as atmospheric deposition and stormwater runoff, for which results were not presented here. As the next step, the database will be developed into an operational state, where national administrations can easily upload and check their data. Further technical development of the database and user interfaces is required, as well as capacity building for data literacy in the involved institutions. Nonetheless, the database is a valuable tool in its current form. It helps establish emission inventories by identifying data gaps that can be filled through monitoring programs. Additionally, it provides a first set of input data for emission modelling using the pathway-oriented approach. Given its usefulness, we have made it available for download at https://doi.org/10.48436/xwve4-h7v43.

Abbreviations

Countries were abbreviated according to ISO norm 3166-1 alpha-2

B[a]P	Benzo[a]pyren (CAS 50-32-8)
BpA	Bisphenol A (CAS 80-05-7)
CAS	Chemical Abstracting Service: Issuing identifiers for chemical substances
DM	Dry matter
DRB	Danube River Basin
EEA	European Environmental Agency
EIONET	European Environment Information and Observation Network
EU	European Union
FAIR	Findable, Accessible, Interoperable, Reusable
ICPDR	International Commission for Protection of the Danube River
JDS	Joint Danube Survey
JRC	EU Joint Research Centre
LOD	Analytical limit of detection
LOQ	Analytical limit of quantification/quantitation
MP	Micropollutants
PAH	Polycyclic aromatic hydrocarbons
PE	Population equivalents
PERMANOVA	Permutational multivariate analysis of variance
PFAS	Per- and polyfluoroalkyl substances
PFBS	Perfluorobutanoic acid (CAS 375-22-4)
PFDA	Perfluorodecanoic acid (CAS 335-76-2)
PFHxA	Perfluorohexanoic acid (CAS 307-24-4)
PFHxS	Perfluorohexane sulfonic acid (CAS 355-46-4)
PFNA	Perfluorononanoic acid (CAS 375-95-1)
PFOA	Perfluorooctanoic acid (CAS 335-67-1)
PFOS	Perfluorooctanesulfonic acid (CAS 1763-23-1)
PFPeA	Perfluoropentanoic acid (CAS 2706-90-3)
PHpA	Perfluoroheptanoic acid (CAS 375-85-9)
PRTR	Pollutant Release and Transfer Register
ROS	Regression on order statistics
RPA	Regionalized pathway analysis
SPM	Suspended particulate matter
TNMN	Danube TransNational Monitoring Network
US-EPA	U.S. Environmental Protection Agency
UWWTD	European Urban Waste Water Treatment Directive
WFD	EU Water Framework Directive
WISE	Water Information System Europe
WWTP	Wastewater treatment plant

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12302-024-00862-4.

Additional file 1. Table S1: Supporting figure 5 - pathway analysis. **Table S2:** Supporting figure 6 - heavy metal concentrations in municipal waste water treatment plant effluents. **Table S3:** Post hoc test for PFAS concentration against plant capacity in municipal waste water treatment plants effluents. **Table S4:** Table supporting figure 7 - PFAS concentrations in municipal waste water treatment plant effluents depending on plant capacity. **Table S5:** Peto - Peto test results for MP soil concentrations against land use. **Table S6:** Supporting figure 8 - MP concentrations in top soil depending on land use.

Acknowledgements

The authors acknowledge the whole Danube Hazard m^{3}c project team, all associated strategic partners and other institutions who supported the work by providing data and other help.

Author contributions

SK: conceptualization, methodology, software, formal analysis, investigation, data curation, writing - original draft, writing - review and editing, visualization. MKa, KD, NW: methodology, software, investigation, data curation, writing - review and editing. AC: conceptualization, methodology, writing - review and editing, supervision, project administration. SP, DS, DKG, AK, DK, CM, MKi, OG: data curation, writing - review and editing. JK: resources, supervision, writing - review and editing. MZ, OZ: conceptualization, methodology, writing - review and editing, supervision, project administration, funding acquisition.

Funding

Open access funding provided by TU Wien (TUW). The data collection was funded by the EU INTERREG Danube Transnational Program in the "Danube Hazard m^{3}c" project with co-funding from the Austrian Federal Ministry of Agriculture, Forestry, Regions and Water Management (BML). Contribution of the MKa and AC has been co-funded by project no. TKP-6-6/PALY-2021, implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the TKP2021-NVA funding scheme. The open access publication was supported by the TU Wien Bibliothek through its Open Access Funding Program.

Availability of data and materials

The database generated and analysed during the current study is available from the TU Wien Research data repository under https://doi.org/https://doi.org/10.48436/xwve4-h7v43 but restrictions apply to a smaller share of the database content, which was used under license for the current study, and so

is not publicly available. In the Additional file tables are available reporting the numbers underlaying the figures presented here (S1, S2, S4, S6) and the results from the applied statistical tests (S3, S5).

Declarations

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]TU Wien, Institute for Water Quality and Resource Management, Karlsplatz 13, 1040 Vienna, Austria. [2]Department of Sanitary and Environmental Engineering, Budapest University of Technology and Economics, Műegyetem Rkp. 3, 1111 Budapest, Hungary. [3]Bulgarian Water Association, Hristo Smirnenski Boulevard 1, 1046 Sofia, Bulgaria. [4]Center for Ecotoxicological Research Podgorica, Bulevar Sarla de Gola 2, 81000 Podgorica, Montenegro. [5]Faculty of Chemical Engineering and Technology, University of Zagreb, Marulićev Trg. 19, 10000 Zagreb, Croatia. [6]International Commission for the Protection of the Danube River (ICPDR), Wagramer Strasse 5, 1220 Vienna, Austria. [7]Department of Environmental Sciences, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia. [8]National Administration Romanian Waters, Edgar Quinet Street 6, 100018 Bucharest, Romania. [9]Water Research Institute, Nábr. Arm. Gen. L. Svobodu 5, 81249 Bratislava, Slovakia. [10]Environment Agency Austria, Spittelauer Lände 5, 1090 Vienna, Austria.

Received: 19 December 2023 Accepted: 11 February 2024
Published online: 09 March 2024

References

1. The European Parliament and the Council of the European Union: 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. Official Journal of the European Union L327:1–72. https://eur-lex.europa.eu/eli/dir/2000/60/oj
2. European Parliament and the Council of the European Union: REGULATION (EC) No 166/2006 concerning the establishment of a European Pollutant Release and Transfer Register and amending Council Directives 91/689/EEC and 96/61/EC. https://data.europa.eu/eli/reg/2006/166/2020-01-01
3. The European Parliament and the Council of the European Union: Directive 2008/105/EC on environmental quality standards in the field of water policy. Official Journal of the European Union. https://data.europa.eu/eli/dir/2008/105/oj
4. The European Parliament and the Council of the European Union: 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. Official Journal of the European Union. https://data.europa.eu/eli/dir/2013/39/oj
5. Anderson MJ (2001) A new method for non-parametric multivariate analysis of variance. Austral Ecol 26:32–46. https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x
6. Anderson MJ (2004) PERMDISP: Permutational analysis of multivariate dispersions. A computer program. University of Auckland, University of Sydney, https://www.yumpu.com/en/document/view/9075657/permdisp-department-of-statistics
7. Anderson MJ (2017) Permutational multivariate analysis of variance (PERMANOVA). In: Balakrishnan N, Colton T, Everitt B, Piegorsch W, Ruggeri F, Teugels JL (eds) Statistics reference online. Wiley, Hoboken, pp 1–15
8. Anderson MJ, Walsh DCI (2013) PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? Ecol Monogr 83:557–574. https://doi.org/10.1890/12-2010.1
9. Anderson MJ, Walsh DCI, Robert Clarke K, Gorley RN, Guerra-Castro E (2017) Some solutions to the multivariate Behrens-Fisher problem for dissimilarity-based analyses. Aust N Z J Stat 59:57–79. https://doi.org/10.1111/anzs.12176
10. Arnold S, Kosztra B, Banko G, Milenov P, Smith, Geoff, Hazeu, Gerard, Bock M, Caetano M, Perger, Christoph, Mancosu, Emanuele (2023) Explanatory Documentation of the EAGLE Concept, Copenhagen. https://land.copernicus.eu/en/technical-library/explanatory-documentation-of-the-eagle-concept-3_2. Accessed 30 Nov 2023
11. Bayerisches Landesamt für Umwelt (LfU): Waterscience service Bavaria: data and information. https://www.gkd.bayern.de/en/. Accessed 30 Nov 2023
12. Bil W, Zeilmaker M, Fragki S, Lijzen J, Verbruggen E, Bokkers B (2021) Risk assessment of per- and polyfluoroalkyl substance mixtures: a relative potency factor approach. Environ Toxicol Chem 40:859–870. https://doi.org/10.1002/etc.4835
13. Björklund K, Bondelind M, Karlsson A, Karlsson D, Sokolova E (2018) Hydrodynamic modelling of the influence of stormwater and combined sewer overflows on receiving water quality: benzo(a)pyrene and copper risks to recreational water. J Environ Manage 207:32–42. https://doi.org/10.1016/j.jenvman.2017.11.014
14. Brielmann H, Döberl G, Weiß S, Grath J (2023) PFAS in Österreichs Grundwasser: Verbreitung, Bewertung und Rolle von Altstandorten als potenzielle Quellen. Österr Wasser- und Abfallw 75:491–502. https://doi.org/10.1007/s00506-023-00976-8
15. Clara M, Strenn B, Kreuzinger N (2004) Carbamazepine as a possible anthropogenic marker in the aquatic environment: investigations on the behaviour of carbamazepine in wastewater treatment and during groundwater infiltration. Water Res 38:947–954. https://doi.org/10.1016/j.watres.2003.10.058
16. COM 540 final (2022). European Commission (EC): Proposal for a Directive of the European Parliament and of the council amending Directive 2000/60/EC establishing a framework for community action in the field of water policy, Directive 2006/118/EC on the protection of groundwater against pollution and deterioration and Directive 2008/105/EC on environmental quality standards in the field of water policy. https://environment.ec.europa.eu/publications/proposal-amending-water-directives_en
17. Conway J, Eddelbuettel D, Nishiyama T, Prayaga SK, Tiffin N (2022) RPostgreSQL: R Interface to the 'PostgreSQL' Database System: R package. https://CRAN.R-project.org/package=RPostgreSQL
18. Cramér H (1946) Mathematical methods of statistics. Princeton mathematical series. Princeton University Press, Princeton
19. Creative Commons: CC BY 4.0 Deed: Attribution 4.0 International. https://creativecommons.org/licenses/by/4.0/. Accessed 28 Nov 2023
20. Dagorn G, Aubert R, Horel S, Martinon L, Steffen T (2023) 'Forever pollution': explore the map of Europe's PFAS contamination. Le Monde, Paris
21. Dowle M, Srinivasan A (2021) Data.table: extension of 'data.frame': R package. https://CRAN.R-project.org/package=data.table
22. Dulio V, van Bavel B, Brorström-Lundén E, Harmsen J, Hollender J, Schlabach M, Slobodnik J, Thomas K, Koschorreck J (2018) Emerging pollutants in the EU: 10 years of NORMAN in support of environmental policies and regulations. Environ Sci Eur 30:5. https://doi.org/10.1186/s12302-018-0135-3
23. Danube hazard m3c: Tackling hazardous substances pollution in the Danube River Basin by measuring, modelling-based management and capacity building. A project in the EU INTERREG Danube Transnational programme. 2020-2023. https://www.interreg-danube.eu/approved-projects/danube-hazard-m3c
24. European Chemicals Agency (ECHA): substance info Card: Mecoprop: EC/List no.: 230-386-8, CAS no.: 7085–19-0. https://echa.europa.eu/de/substance-information/-/substanceinfo/100.027.624. Accessed Nov 2023
25. European Commission, Directorate-General for Environment (2012): Technical guidance on the preparation of an inventory of emissions, discharges and losses of priority and priority hazardous substances. Guidance document No 28.
26. European Environment Agency (EEA): mercury in Europe's environment: a priority for European and global action. EEA report, 11/2018. Publications Office of the European Union, Luxembourg

27. European Environment Agency (EEA) (2019): Waterbase— UWWTD: Urban Waste Water Treatment Directive—reported data: Version 6. https://www.eea.europa.eu/data-and-maps/data/waterbase-uwwtd-urban-waste-water-treatment-directive-6. Accessed 28 Nov 2023
28. European Environment Agency (EEA) (2022): EIONET Data dictionary: Vocabulary: Observed property. WISE—Water Information System for Europe. http://dd.eionet.europa.eu/vocabulary/wise/ObservedProperty/
29. European Environment Agency (EEA) (2023): European Industrial Emissions Portal: Data from the reporting under the PRTR and the IE directive. https://industry.eea.europa.eu/about
30. European Environment Agency (EEA): waterbase—water quality ICM. https://sdi.eea.europa.eu/catalogue/srv/api/records/fbf3717c-cd7b-4785-933a-d0cf510542e1. Accessed 28 Nov 2023
31. European Commission (EC): Commission Implementing Decision (EU) 2020/1161 of 4 August 2020 establishing a watch list of substances for Union-wide monitoring in the field of water policy pursuant to Directive 2008/105/EC of the European Parliament and of the Council. Official Journal of the European Union. http://data.europa.eu/eli/dec_impl/2020/1161/oj
32. Fuchs S, Scherer U, Wander R, Behrendt H, Venohr M, Opitz D, Hillenbrand T, Marscheider-Weidemann F, Götz T (2010) Calculation of Emissions into Rivers in Germany using the MONERIS Model: Nutrients, heavy metals and polycyclic aromatic hydrocarbons. Federal Environment Agency, Dessau-Roßlau
33. Fuchs S, Weber T, Wander R, Toshovski S, Kittlaus S, Reid L, Bach M, Klement L, Hillenbrand T, Tettenborn F (2017) Effizienz von Maßnahmen zur Reduktion von Stoffeinträgen. Endbericht, Umweltbundesamt, Dessau-Roßlau
34. Fuchs S, Kaiser M, Kiemle L, Kittlaus S, Rothvoß S, Toshovski S, Wagner A, Wander R, Weber T, Ziegler S (2017) Modeling of regionalized emissions (MoRE) into water bodies: an open-source river basin management system. Water 9:239. https://doi.org/10.3390/w9040239
35. Gawlik BM (2023) JRC FATE monitoring database on occurrence and levels of chemical contaminants. https://ipchem.jrc.ec.europa.eu/#showmetadata/FATE
36. Gomez Cortes L, Marinov D, Sanseverino I, Navarro Cuenca A, Niegowska M, Porcel Rodriguez E, Stefanelli F, Lettieri T (2022) Selection of substances for the 4th watch list under the water framework directive. Publications Office of the European Union, Luxembourg
37. Helsel DR (2006) Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it. Chemosphere 65:2434–2439. https://doi.org/10.1016/j.chemosphere.2006.04.051
38. Helsel DR (2011) Statistics for censored environmental data using minitab® and R, 2nd edn. Hoboken, Wiley
39. International Commission for the Protection of the Danube River (ICPDR) (2021): Danube river basin management plan, 2021st edn. ICPDR, Vienna
40. International Commission for the Protection of the Danube River (ICPDR) (2023): Danube river basin water quality database. https://wq-db.icpdr.org
41. Isaacs KK, Wall JT, Williams AR, Hobbie KA, Sobus JR, Ulrich E, Lyons D, Dionisio KL, Williams AJ, Grulke C, Foster CA, McCoy J, Bevington C (2022) A harmonized chemical monitoring database for support of exposure assessments. Sci Data 9:314. https://doi.org/10.1038/s41597-022-01365-8
42. IUSS Working Group WRB (2022): World reference base for soil resources: International soil classification system for naming soils and creating legends for soil maps, 4th edn. Vienna, International Union of Soil Sciences (IUSS)
43. Julian P, Helsel DR (2023) NADA2: Data Analysis for Censored Environmental Data: R package. https://cran.r-project.org/web/packages/NADA2/index.html
44. Kitamura S, Suzuki T, Sanoh S, Kohta R, Jinno N, Sugihara K, Yoshihara S, Fujimoto N, Watanabe H, Ohta S (2005) Comparative study of the endocrine-disrupting activity of bisphenol A and 19 related compounds. Toxicol Sci 84:249–259. https://doi.org/10.1093/toxsci/kfi074
45. Kittlaus S, Clara M, van Gils J, Gabriel O, Broer MB, Hochedlinger G, Trautvetter H, Hepp G, Krampe J, Zessner M, Zoboli O (2022) Coupling a pathway-oriented approach with tailor-made monitoring as key to well-performing regionalized modelling of PFAS emissions and river concentrations. Sci Total Environ 849:157764. https://doi.org/10.1016/j.scitotenv.2022.157764
46. Kosztra B, Büttner G, Hazeu G, Arnold S (2019) Updated CLC illustrated nomenclature guidelines. Service Contract No 3436/R0-Copernicus/EEA.57441 Task 3, D3.1—Part 1., Vienna, Austria. https://land.copernicus.eu/en/technical-library/clc-illustrated-nomenclature-guidelines. Accessed 28 Nov 2023
47. Lamprea K, Bressy A, Mirande-Bret C, Caupos E, Gromaire M-C (2018) Alkylphenol and bisphenol A contamination of urban runoff: an evaluation of the emission potentials of various construction materials and automotive supplies. Environ Sci Pollut Res 25:21887–21900. https://doi.org/10.1007/s11356-018-2272-z
48. Lee L (2020) NADA: nondetects and data analysis for environmental data: R package. https://CRAN.R-project.org/package=NADA
49. Liska I, Wagner F, Sengl M, Deutsch K, Slobodník J, Paunovic M (2021) Joint Danube survey 4 scientific report: a shared analysis of the Danube river. JDS4, Wien
50. Lonappan L, Brar SK, Das RK, Verma M, Surampalli RY (2016) Diclofenac and its transformation products: environmental occurrence and toxicity—a review. Environ Int 96:127–138. https://doi.org/10.1016/j.envint.2016.09.014
51. Mangiafico SS (2023) rcompanion: Functions to Support Extension Education Program Evaluation, New Brunswick, New Jersey. https://CRAN.R-project.org/package=rcompanion/
52. Müller V, Kindness A, Feldmann J (2023) Fluorine mass balance analysis of PFAS in communal waters at a wastewater plant from Austria. Water Res 244:120501. https://doi.org/10.1016/j.watres.2023.120501
53. Nickel JP, Sacher F, Fuchs S (2021) Dataset of micropollutant concentrations and standard water quality parameters in wastewater treatment plants, combined sewer overflows, and stormwater outfalls in Germany. https://doi.org/10.35097/449
54. Oksanen J, Simpson GL, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Solymos P, Stevens MHH, Szoecs E, Wagner H, Barbour M, Bedward M, Bolker B, Borcard D, Carvalho G, Chirico M, Caceres MD, Durand S, Evangelista HBA, FitzJohn R, Friendly M, Furneaux B, Hannigan G, Hill MO, Lahti L, McGlinn D, Ouellette M-H, Cunha ER, Smith T, Stier A, Braak CJFT, Weedon J (2022) Vegan: community ecology package: R package. https://CRAN.R-project.org/package=vegan
55. Pistocchi A, Cinnirella S, Mouratidis P, Rosenstock N, Whalley C, Sponar M, Pirrone N (2022) Screening of Mercury pollution sources to European inland waters using high resolution earth surface data. Front Environ Sci. https://doi.org/10.3389/fenvs.2022.1021777
56. PostgreSQL Global Development Group (1996–2023) PostgreSQL. https://www.postgresql.org/about/
57. R Core Team (2022) R: a language and environment for statistical computing, Vienna, Austria. https://www.R-project.org/
58. Schauberger P, Walker A (2023) Openxlsx: read, write and edit xlsx files: R package. https://CRAN.R-project.org/package=openxlsx
59. Schreiberová M, Vlasáková L, Vlček O, Šmejdířová J, Horálek J, Bieser J (2020) Benzo[a]pyrene in the ambient air in the Czech Republic: emission sources, current and long-term monitoring analysis and human exposure. Atmosphere 11:955. https://doi.org/10.3390/atmos11090955
60. Shareef A, Angove MJ, Wells JD, Johnson BB (2006) Aqueous solubilities of estrone, 17β-estradiol, 17α-ethynylestradiol, and bisphenol A. J Chem Eng Data 51:879–881. https://doi.org/10.1021/je050318c
61. Slobodnik J, von der Ohe PC (2015) Identification of the Danube river basin specific pollutants and their retrospective risk assessment. In: Liska I (ed) The Danube river basin. Springer, Berlin, pp 95–110
62. Slobodník J (2023) EMPODAT database: database of geo-referenced monitoring data on emerging substances in air, water and soil. https://www.norman-network.com/nds/empodat/. Accessed 07 Nov 2023
63. Weller P, Popovici M (2012) Danube river basin management—rationale and results: how to link science, as the basis for Policy. River Systems 20:103–109. https://doi.org/10.1127/1868-5749/2011/020-0034
64. Whalley C, Mohaupt V, Busch W, van den Roovart J, van Dujnhofen N, Kirst I, Schmedtje U, Altenburger R, Sommer L (2018) Chemicals in European waters: Knowledge developments. EEA report, no. 18/2018. Publications Office of the European Union, Luxembourg
65. Wickham H, Ooms J, Müller K (2022) RPostgres: Rcpp Interface to PostgreSQL: R package. https://CRAN.R-project.org/package=RPostgres

66. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, Da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3:160018. https://doi.org/10.1038/sdata.2016.18

Publisher's Note