Joint Research Programme
BTO 2023.039 | August 2023

Environmental forensics, signatures of pollution

Joint Research Programme

KWR

Bridging Science to Practice

Report

Environmental Forensics, signatures of pollution

BTO 2023.039| August 2023

This research is part of the Joint Research Programme of KWR, the water utilities and Vewin.

Project number
402045/296

Project manager
Patrick Bäuerlein

Client
BTO - Thematical research - Chemical safety

Author(s)
Tessa Pronk, Elvio Amato

Quality Assurance
Thomas ter Laak

Sent to
This report is distributed to BTO-participants.

This report will become public one year after publication.

Keywords

Clustering; patterns; chemical pollution; monitoring; water quality; emission

Year of publishing
2023

More information
Dr. ir. Tessa Pronk
T +31 6 29337806
E Tessa.Pronk@kwrwater.nl

PO Box 1072
3430 BB Nieuwegein
The Netherlands

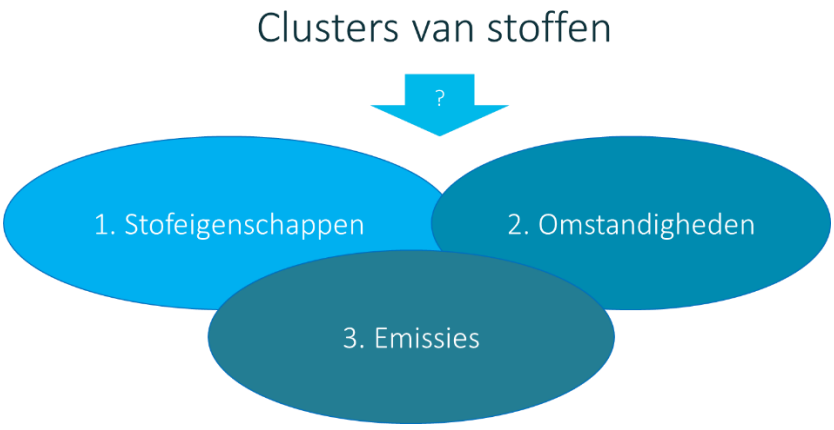
T +31 (0)30 60 69 511
E info@kwrwater.nl
I www.kwrwater.nl

Managementsamenvatting

Environmental Forensics: stoffen die gelijk variëren in concentratie linken aan stofeigenschappen, omstandigheden en mogelijke emissies, geeft inzicht in vervuiling

Auteur(s) Dr. ir. Tessa Pronk, dr. Elvio Amato

In de Nederlandse oppervlaktewateren komen veel verschillende stoffen voor. Metingen tonen dat hun concentraties variëren. Het is vaak onduidelijk waarom een bepaalde stof op een bepaald moment op een bepaalde plek in een hoge concentratie wordt gemeten. Gebeurtenissen reconstrueren die leiden tot verhoogde concentraties kan met technieken uit het vakgebied '*Environmental Forensics*'. In dit rapport is gekeken naar concentraties van stoffen in samenhang. Met clustering-technieken zijn voor verschillende locaties in Rijn en Maas groepen van stoffen ('clusters') vastgesteld. In deze clusters variëren stoffen in metingen op eenzelfde manier in concentratie. Door de stoffen als cluster te linken met 1) stofeigenschappen 2) gemeten omstandigheden 3) aanwezigheid van stoffen die emissiebronnen, stoftypen, of type gebruik vertegenwoordigen in het cluster, kunnen hypothesen worden geformuleerd rond de oorzaak van de clustering. Door op deze manier clusters te interpreteren kunnen drinkwaterbedrijven de concentraties van stoffen op een bepaalde locatie doorgronden en dat geeft de mogelijkheid hierop te anticiperen. Met de hypothesen rond gevonden clusters kan nader onderzoek worden gedaan en kunnen zo nodig verantwoordelijke partijen tot maatregelen worden gemaand.



Overzicht van de drie factoren zoals geanalyseerd in dit rapport die gelinkt worden aan groepen van stoffen ('clusters') met gelijke variaties in concentratie in oppervlaktewateren.

Belang: Anticiperen op- of aanpakken van vervuiling

Oppervlaktewateren in Nederland bevatten veel verschillende stoffen. Deze worden regelmatig gemeten en worden daarbij in verschillende concentraties aangetroffen. Het is vaak onduidelijk waarom een bepaalde stof op een bepaald moment en op een bepaalde plek een in een hoge concentratie wordt gemeten. Is het een gevolg van een recente lozing of zorgen tijdelijke omstandigheden voor een hoge concentratie? En waarom worden dan slechts specifieke stoffen aangetroffen? Waterbedrijven hebben behoefte aan meer inzicht in de oorzaken van deze variërende concentraties om in de toekomst te kunnen anticiperen op verwachte concentraties en daarnaast noodzakelijke maatregelen te formuleren en uit te voeren.

Aanpak: Clusters linken aan stofeigenschappen, omstandigheden en referentie-stoflijsten

Met technieken uit het vakgebied '*Environmental forensics*' is een methode opgezet om nader te kunnen kijken naar concentraties van stoffen in samenhang. Met clustering-technieken zijn voor verschillende locaties in Rijn en Maas groepen van stoffen gemaakt op basis van historische concentratiedata (2017-2021) van oppervlaktewaterlocaties in de Rijn en de Maas. Binnen deze clusters variëren de stoffen op eenzelfde manier in concentratie. Door de stoffen als cluster te vergelijken met stofeigenschappen zoals oplosbaarheid, omstandigheden zoals regenval en referentielijsten van stoffen die wijzen op een bepaalde emissieoorzaak (bijvoorbeeld stoffen die gebruikt worden in aardappelteelt of stoffen die als biocide in bouwmateriaal worden gebruikt) kan een hypothese worden geformuleerd rond de oorzaak van de concentraties van stoffen in het cluster.

Resultaten: 180 clusters op 18 locaties in Rijn en Maas

Er werden meerdere clusters per locatie gevonden: in totaal 180 clusters op 18 locaties in Rijn en Maas. Sommige clusters waren uniek voor een locatie. Ook waren clusters af en toe geïsoleerde 'incidenten' waarin een aantal stoffen in een enkele meting plotseling hoog waren. Een achttal verschillende clusters werden op meerdere locaties in ongeveer dezelfde samenstelling gevonden. Deze 'herhalende clusters' zijn in meer detail geanalyseerd om oorzaken te achterhalen. Een herhalend cluster met met PCB's werd bijvoorbeeld gekenmerkt door stoffen met de eigenschappen hoge persistentie en lage mobiliteit en was daarnaast gelinkt met hoge waterstanden en een hoog zuurstofgehalte: dit kan mogelijk duiden op resuspensie uit slib. Een cluster met farmaceutica was afhankelijk van rivierafvoer, dat zorgde voor verdunning dan wel concentratie van de stoffen. Niet alle stoffen werden in clusters geplaatst. Hun concentratie varieerde niet op eenzelfde manier als die van andere stoffen.

Toepassing: Gedetailleerde interpretatie per locatie

Met de resultaten uit dit project kan in vervolgonderzoek verder tot in detail bekeken worden hoe de gevonden clusters kunnen worden geïnterpreteerd. Met de vervolgens te formuleren hypothesen rond oorzaken kan nader onderzoek gedaan worden en kunnen zo nodig verantwoordelijke partijen tot maatregelen worden gemaand. De resultaten vergroten nu al het begrip rond gevonden concentraties van stoffen en de mogelijkheid hierop te anticiperen. De methoden die ontwikkeld zijn kunnen ook op andere, meer lokaal beïnvloede oppervlaktewateren worden toegepast wanneer een historische meetreeks uit een uitgebreid monitoringsprogramma beschikbaar is.

Rapport

Dit onderzoek is beschreven in het rapport *Environmental Forensics, signatures of pollution* (BTO 2023.039)

Bijbehorend datapakket:

- https://doi.org/10.5281/zenodo.8220952

Via een 'R Shiny-app' kan de associatie van een individuele stof met de gevonden clusters opgezocht worden. Bijbehorende applicatie:

- https://tessaepronk.shinyapps.io/ShinyEnvFor/

Andere relevante rapporten:

BTO 2016.105 Verbetering prognose waterkwaliteit bij innamepunten van oppervlaktewater voor de drinkwatervoorziening: https://library.kwrwater.nl/publication/55745874/

Peer reviewed publicatie:

Pronk et al., 2024 Linking Clusters of Micropollutants in Surface Water to Emission Sources, Environmental Conditions, and Substance Properties: https://doi.org/10.3390/environments11030046

Meer informatie
Dr. ir. Tessa Pronk
T 030-6069681
E tessa.pronk@kwrwater.nl

PO Box 1072
3430 BB Nieuwegein
The Netherlands

Contents

1 Introduction

The presence of anthropogenic substances in freshwater sources poses a challenge to the drinking water sector. Information on the sources and emission routes of contaminants in surface water and groundwater is often lacking, which makes it difficult to define measures to reduce or prevent emissions. However, environmental forensic approaches have shown to be useful for linking chemicals to sources of contamination. In broad terms, environmental forensics involves the reconstruction of the chain of events that lead to episodes of contamination in the environment. Investigations typically aim at understanding the links between contamination sources and release in the environment, and in some cases may also involve the establishment of the legal responsibility of the contamination event in a regulatory context. Typical investigation techniques include chemical fingerprinting, chemical fate and transport modelling, hydrogeological investigation, and reconstructing operational histories. These techniques have been applied in many different scenarios, including urban and remote areas, and using varying environmental matrixes (e.g., water, sediment, soil, biota). For instance, the occurrence of one or multiple chemicals with respect to their background concentrations, spatial and temporal distribution, can be used to reconstruct contamination events (Warner et al., 2019; Yang et al., 2020). Or, chemical indicators can be used to investigate connectivity of waterbodies, e.g., between urban, agriculture and natural environments (Pascual-Aguilar et al., 2013). In Appendix VII a more extensive literature review on techniques can be found.

Compliance with water quality regulations requires extensive monitoring campaigns to be carried out, which result in large datasets of chemical measurements. Since the list of monitored substances is constantly growing, the size of monitoring datasets also consistently increases. The use of these datasets is generally limited to the comparison with water quality guidelines, and, in case of major contamination episodes or calamities, to applications for forensic investigations. In contrast, little effort is dedicated to mining of underlying information that may likely exist in such datasets. For instance, statistical techniques that detect specific signatures, associations, and co-occurrence of substances can be used to investigate patterns and relationships between substances in such datasets, and contribute to revealing underlying information that is not immediately evident, and thus often overlooked.

In this study, we investigate the aspect of 'chemical context' for applications in environmental forensics investigations by clustering analysis. The co-occurrence of substances in monitoring data can be exploited to create associations with specific sources of pollution. Similarly to words in a text, the meaning of the presence of a substance may be interpreted in the context of other substances in the same measurements. To give a hypothetical example; permethrin is commonly used as an insecticide for wood preservation treatments (Arip et al., 2013) and also as an antiparasitic for veterinary medicine (Carabajal et al., 2021). If at a location more wood preservatives are found with permethrin, it can be inferred that this location is affected by activities related to wood treatment or construction activities. In contrast, if other veterinary-related substances are found, the occurrence of permethrin may indicate emissions linked to animal farming. In other words, the context of measured substances can assist with interpretating the occurrence of individual substances. Moreover, the cooccurrence of substances can potentially be linked to substance properties that substances have in common, and/or common environmental circumstances that are associated with the found substances.

With regard to the analyses, firstly statistical methods are performed to identify groups of substances that frequently occur together in large scale datasets – i.e. clusters – using large sets of structural monitoring data of surface water over several years. Consequently it is investigated 1) How clustered measurement data can be interpreted by looking at the presence substances in 'reference lists' of substances that are linked to specific emission sources (such as an industry type) or emission causes (such as use of insecticides) 2) If specific substance properties can be linked to the identified clusters. 3) If specific environmental conditions can be linked to the clusters. Then, it is possible to form a hypothesis on why the substances occur in the identified cluster.

2 Methods

2.1 Environmental monitoring data

The basis for the analyses are monitoring data from RIWA-Rijn and RIWA-Maas (Table 1). This data contains concentrations of a wide range of contaminants in surface water for a spectrum of locations along the river Rhine and Meuse. We consider data for the years 2017-2021. Parameters are, for the most part, measured every 4 weeks.

Table 1. Data used for exploratory analyses

	RIWA-Rijn	RIWA-Maas
Temporal spread	2551 unique sampling events over 5 years (by date)	2323 unique sampling events over 5 years (by date)
Spatial spread	9 locations	13 locations
Monthly aggregated data for clustering	539 samples, 854 substances (with a CAS-number)	646 samples, 1008 substances (with a CAS-number)
Weekly aggregated data for clustering	1128 samples, 854 substances (with a CAS-number)	2315 samples, 1008 substances (with a CAS-number)

The RIWA datasets contain measurements for a large number of substances. In many measurements substances occur at concentrations below the reporting limit (RL) and these were indicated in the dataset using the symbol "<". Concentrations measured <RL were replaced by zeros. This helps to recognise these values later and remove all substances that were never measured above RL. Not all substances were measured in all samples (i.e., not all substances were always included in the method used to analyse samples). As a result, the datasets are populated by many 'missing values', which indicate that a given substance is not measured in a sample. This is a problem for clustering algorithms. Aggregation to weekly measurements was performed to create a more complete dataset. Weekly aggregation was chosen because enough samples and substances remained after weekly aggregation, and aggregation by month has some disadvantages. Namely, combining measurements per month will result in parameters appearing as a single measurement while these were not necessarily measured in the same sample or at the same date. Also, if multiple measurements are done in a month, these need to be condensed into a single value by taking for instance the average value. Weekly aggregated data had, on average, four extra samples per location compared to monthly aggregated data. The number of substances per location decreased on average 5% and this was worst for the locations with less measured parameters (up to 19% loss). Nevertheless, these results indicate that the monthly measured substances are for the *most* part all measured within the timeframe of the same week. Because of the increased accuracy of weekly aggregated data, the weekly aggregated data was chosen for further analysis.

Figure 1. Meuse and Rhine locations with water quality measurement locations, used in this project (also see Table 2).

2.2 Identifying clusters in environmental monitoring data

Cluster analysis (CA) or clustering is a collection of different methods to group observations in such a way that observations in the same group (a 'cluster') are more similar to each other than to those in other groups (clusters). In our case, observations were the concentrations of parameters. For finding clusters (which we can also call 'signatures', or 'fingerprints') of parameters having similar concentration patterns over the samples in monitoring data, we perform a hierarchical clustering. Hierarchical clustering (also called hierarchical cluster analysis or HCA) is a specific method of cluster analysis which seeks to build a hierarchy of clusters. There is no prior information on group membership needed for the clustering. Only the values for the observations are used to compute a measure of similarity. The clustering can be performed both on the parameters (i.e. the chemicals), as on the samples (measured on different locations at different times). In a HCA, dendrograms that illustrate a hierarchy are used to show relationships between parameters. Parameters that fall into another cluster only towards the bottom of the hierarchy are more similar than parameter that fall into other clusters higher up in the hierarchy (see Figure 2). The vertical length of the branches is an indicator for similarity, the shorter the branches the more similar the parameters are.

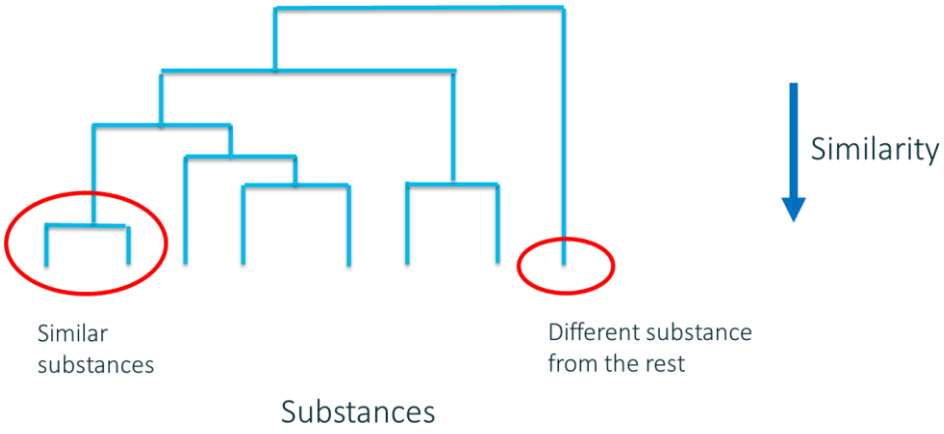


Figure 2. Dendrogram example. Splits at the top of the dendrogram indicate (in this report) large differences in substance concentration patterns. Splits towards the bottom indicate greater similarities.

2.2.1 Distinguishing water sample types

Prior to determining clusters of substances in historical data with HCA or any other CA method, it is important to see if there are different sample types. For the RIWA data, samples may differ based on the location and sampling date (i.e., season or year). This is important for the identification of clusters. Patterns can be disrupted if substances are emitted by one particular source (with accompanying substances) in one sample type, and by another source (with accompanying substances) in the other sample type. If the sample types are not distinguished and analysed separately, this will weaken the correlation between substances and will fail to uncover the correlation that is only present in one type of samples.

Determining if there are different sample types can be achieved by unsupervised clustering. We investigated if the samples of the RIWA dataset were very different between years, seasons, or locations. We perform the analysis by a principal component analysis (PCA). PCA is a statistical procedure that summarizes the information in large data tables by a smaller set of “summary indices” that can be more easily visualized (as points in a 2D or 3D plot) and analyzed. The points (representing the information in samples) in the 2D or 3D plot can be colored according to the factors years, months or locations to see if any groups have appeared in the PCA that corresponds to these factors. If we know sample types can be distinguished and split the data accordingly before analyses, we know that we have a good chance to find patterns of structurally co-occurring substances with HCA.

2.2.2 Pre-processing data for clustering

No missing values are permitted in clustering. Because in our case there is no accurate way to replace values with estimated values, especially with possibly highly varying concentration data, we choose to remove missing values. To efficiently remove missing values while maintaining as much data as possible, an algorithm was applied that automatically removes either the parameter or a sample with relatively high fraction of missing values. This is repeated until the dataset no longer contains missing values. Locations that had only a few samples (<20) with measurement data were omitted from the analyses. Additionally, parameters without *any* measurements above the reporting limit were removed.

The concentrations of the substances in a sample influence the result of the clustering. By scaling the data, the actual height of the concentrations will not influence the results and only the relative concentration will induce a difference. To achieve this, data per parameter was scaled to a ‘Z-score’. The Z-score is the number of standard deviations a given data point lies from the mean. For data points that are below the mean, the Z-score is negative. The formula for calculating the Z-score is $z = (x-\mu)/\sigma$, where x is the concentration of a given substance, μ its mean concentration, and σ its standard deviation. Typically, Z-score values are between -3 and 3. This small interval makes the influence

of different substances more comparable. Prior to clustering, samples that had overall high Z-scores for the substances were excluded as 'outliers'. This is because regular patterns in the clusters can be disrupted by the influence of such deviating samples. This was done based on visual inspection. These samples are separate from the other samples in a dendrogram (e.g. see Figure 2, a 'substance' would be a 'sample').

In order to generate clusters, an appropriate metric (a measure of distance between pairs of observations/ parameters) and a linkage criterion which specifies the dissimilarity of sets as a function of the pairwise distances of observations/parameters in the sets is used. For the analysis in this report, 'Euclidean distance' was used as a similarity measure in the HCA, to estimate the distance or (dis)similarity of each pair of observations. 'Ward' was used as the linking criterion. This manner of clustering minimizes within-cluster variance. This is in agreement to used settings in genomics analyses, another field where clustering is often used to find signatures and genetic resemblance.

Unfortunately, there is no such thing as the objectively best clustering. Different methods are better or less suited to bring different patterns to the surface. These settings produced visually concise clusters for the various locations and therefore were chosen as the clustering settings of preference.

2.2.3 Assigning Cluster significance

For our purpose to select clusters of substances that change in concentration together, we want to only select clusters that look consistent in a heatmap (e.g. Figure 8) and are consistent throughout the dendrogram towards the bottom where these would in a random situation increasingly split up in separate clusters (e.g. Figure 2). To determine such clusters a 'cluster significance' (ClusSig) method was developed. More details can be read in Appendix III.

In short, the ClusSig method works with the simple assumption that any relatively large found cluster compared to an expected *randomly occurring* size is extraordinary and points to a real (not random) cluster. It works as follows. At every level of the hierarchy a distribution of randomly drawn cluster sizes is simulated. Some cluster sizes will be very rare and others common. At a determined level in the cluster hierarchy (where the overall significance of substances is highest) the size of the real, actual cluster sizes in the data are compared to the random distribution. If the size of any cluster at this level is very rare (less than 10 percent of simulated random clusters have this size) it is considered significant.

2.3 Linking extra information clusters of substances

Once clusters are established, other information is linked to these clusters of substances. Three types of information are considered (see Figure 5) and this is explained below.

2.3.1 Associating emissions via reference lists of substances to clusters.

If the substances in the cluster consist of substances that are typically associated with one specific use, source or other origin, this provides a hypothesis on why the substances are found as a cluster. For instance, cluster could consist of substances that are typically used in potato cultivation. This could mean that the cluster is caused by pest control in potato cultivation. In this report we refer to lists of substances with a commonality in use or origin as 'Reference lists'. Reference lists were compiled using literature data and (public) lists of chemicals (Table 2). For an overview, see the Table in Appendix I. For all chemicals in all Reference lists, see the data package (https://doi.org/10.5281/zenodo.8220952, 2023).

The sources for these lists included: chemicals used in different agricultural activities in certain quantities for the production of fruit, chemicals used in agriculture for the production of vegetables (Dutch Central bureau of statistics,

CBS), chemicals measured in effluent of Dutch sewage treatment plants (Watson database), chemicals measured in close proximity of different agricultural activities ("Landelijk meetnet Gewasbeschermingsmiddelen", LM GBM), chemicals consistently measured in the trans-border rivers Meuse and Rhine (database RIWAbase), biocide product types (ECHA website), parameter groups used in the database of substances of RIWA-Rijn 2022, a review of substances and emissions published by Warner et al. (2019), substances listed in the European Environmental Agency (EEA) industrial emissions database (limited to emissions in water), a list of veterinary pharmaceuticals found in manure (Rakonjac et al., 2022) or illicit drugs (RIVM, 2022) and several lists such as 'veterinary pharmaceuticals' as listed in the 'Comptox chemical lists'. In these lists a total of 1968 unique substances are included.

Most substances are relatively unique, present in one reference list, while others are associated with many different reference lists. The list that contains many substances is 'Crop sectors, total' (170). Substances that occur in relatively many lists are Glyphosate (29), Thiacloprid (19), Deltamethrin (18), Azoxystrobin (18). This has partly to do with the fact that these chemicals are applied, both in agriculture (crop protection), in household setting (gardening, pest control) and by municipalities (remove weeds, control insects) which leads to membership in different Reference lists. Overlapping lists were merged (see Appendix I for a more in depth explanation). This resulted in 164 separate reference substance lists. The overall similarity of the Reference lists (expressed as the % remaining overlap in substances) is visualised as a hierarchical clustering in Appendix I. Some lists are still more related than others.

For matching substances found in clusters with substances present in Reference lists we use a 'hypergeometric test'. This method tests significant overlap ('enrichment') of two lists. Any two lists can be compared, resulting in a *p*-value for significance of the overlap (see Figure 3). This technique is frequently used in genomics research, to link gene expression patterns to known gene expression pathways (Hermsen et al., 2013). The information that is required as input for this method is:

- M, the total number of relevant chemicals (in all reference lists and monitoring data)
- n, the chemicals in a reference substance list
- N, the number of chemicals found above the reporting limit in a cluster of monitoring substances (or a monitoring sample)
- X, the size of the overlap

We can then compute a probability of drawing X chemicals out of N from a measurement containing n reference chemicals out of all chemicals M in the following way: p-value = hypergeometric test (x-1, M, n, N).

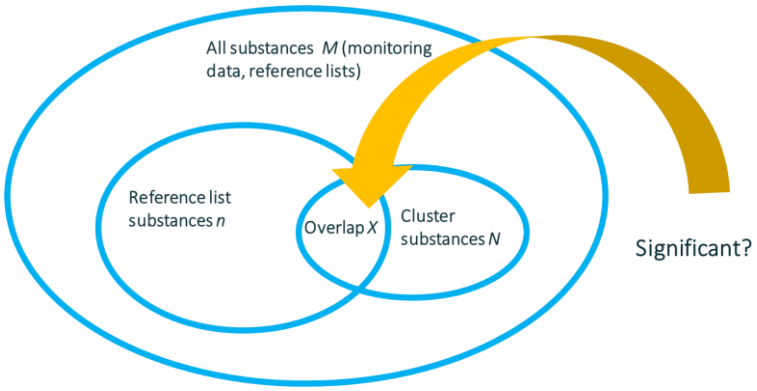


Figure 3. Visualisation of the hypergeometric test for significance of enrichment of Reference list substances in clusters. If the overlap (X) between a given Reference list and a given cluster is larger than a random expected overlap, a low p-value is generated (indicating significant overlap).

If two lists N and n are very small compared to the total number of relevant chemicals M, it is arguably rare that these two have overlapping substances. A small overlap X will be significant. In contrast, if both lists n and N are big, the overlap is to be expected. Only a big overlap X will be significant. The significant score is considered and reported in the current analyses only if two or more substances overlap between the measured data in a sample and the reference list.

2.3.2 Associating substance properties to found clusters

To test if any detected clusters can be linked to substance properties, several substance property values were calculated by (open source) models. These were 'EpiSuite' models via the EpiSuite software and 'Opera' models via the Batch mode in the CompTox Chemical Dashboard. Inorganic substances were not suitable for these calculations, and therefore excluded from the analysis. We made boxplots of these substance property values, per cluster (see Appendix IV).

2.3.3 Associating environmental conditions to found clusters

The influence of environmental conditions on clusters (inclusive inorganic substances) was analysed. These included conditions measured in the RIWA-Rijn and RIWA-Maas datasets, such us oxygen, river discharge, pH, dissolved organic content (DOC) and temperature. Additional conditions that were included like precipitation, sunny hours per day and evaporation potential. These daily measurements were downloaded from the Dutch knowledge for weather, climate, and seismology KNMI. Each weather station was linked to the closest monitoring location in Rhine and Meuse and data were aggregated per week by taking the mean.

The influence of environmental conditions on clusters was evaluated in a similar way to that of substance properties. This resulted in a rather complex analysis to link environmental conditions to substance concentrations. First, it was identified in what samples substances had relatively high (top ten percent) concentration values. This was done for Rhine and Meuse separately. If the concentration of a substance was structurally higher in one location within the same river system, a correction by normalisation of the concentration values was applied. This lead to an equal chance for locations to contribute to relatively high concentration values, and all locations could be represented. The value of the environmental condition was administrated for samples in that high ten percent of the substance concentrations. Then, per substance, the mean was taken of the 'condition' in the samples where the substance had those high concentrations. This resulted in a mean value of the condition that is associated with high concentrations of a substance. We made boxplots of these values of conditions per cluster, and Meuse and Rhine separately (see Appendix IV).

3 Results: Statistics

3.1 Sample types in environmental monitoring data

The first analysis focused on the identification of specific sample types. In Figure 4 we show results of unsupervised clustering of four locations of the RIWA-Rijn data (aggregated per month, missing values removed) with a 3D PCA plot. The samples in Figure 4 are colored by location (left) and years (right). In Figure 4 it can be observed that the clusters are highly defined by location. This observation remains if the plot is made for all locations of the RIWA-Rijn data (not shown) and also the RIWA-Maas data (not shown). If the samples are colored by season, no obvious clustering appears (data not shown). This shows that we can take the location as a starting point to derive clusters.

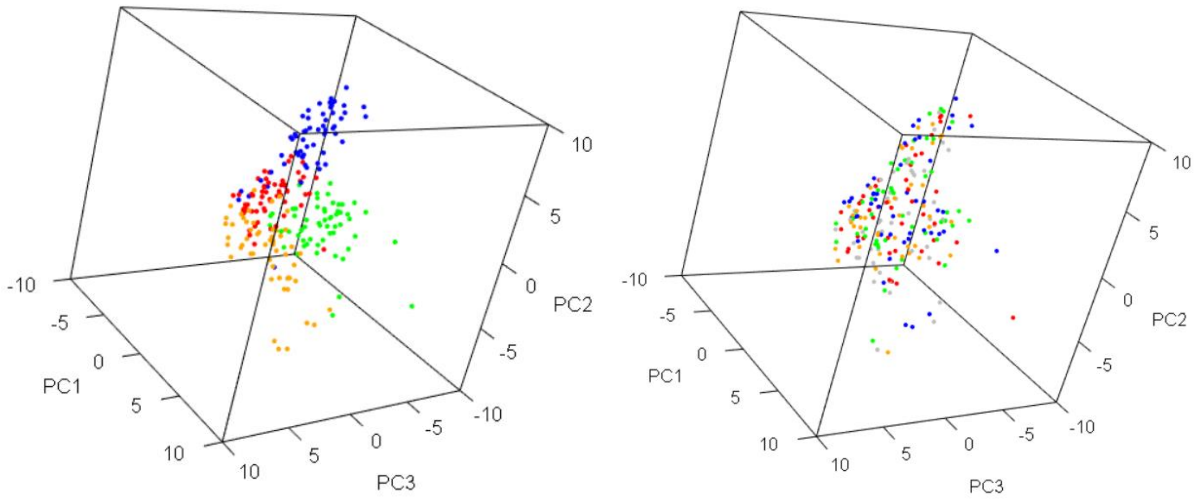


Figure 4. Similarity between RIWA-Rijn time/place samples with a PCA score. Samples from the same place are alike (left). In contrast, the period (year) has no noticeable influence on clustering (right). Colors in the left plot refer to Andijk (blue), Nieuwersluis (green), Nieuwegein (red) and Lobith (orange). The colors in the right plot refer to years 2017-2021.

3.2 Clusters in Meuse and Rhine locations

Relevant clusters were identified with the ClusSig approach. There are about ten 'significant' clusters per location. The average size of the clusters is 6 substances. There is a clear positive link between the number of substances in a location and the number of clusters identified. Also, the more clusters, the smaller on average the cluster size (from 6.5 to 4.5 substances). This may indicate that the 'optimal level of cluster number' is chosen more towards the bottom of the hierarchy when a lot of substances are involved. Determining the 'optimal level' may have to be reevaluated in future applications.

Table 2. Overview of clusters per location, identified with the ClusSig approach.

Location code	Location name	Year / months / week	Substances	Clusters	Substances in clusters	Average cluster size	River
AND	Andijk	62	168	13	64	4.9	Rhine
LOB	Lobith	52	193	18	100	5.6	Rhine
NGN	Nieuwegein	63	201	22	102	4.6	Rhine
NSL	Nieuwersluis	64	139	10	68	6.8	Rhine
BRI	Brienenoord	62	121	8	52	6.5	Rhine
KAM	Kampen	64	109	10	53	5.3	Rhine
KMW	Ketelmeer-West	61	102	10	59	5.9	Rhine
MMM	Markermeer-Midden	60	94	8	51	6.4	Rhine

VWZ	Vrouwezand (IJsselmeer)	61	88	6	37	6.2	Rhine
BRA	Brakel	53	164	16	75	4.7	Meuse
HEE	Heel	52	163	18	76	4.2	Meuse
EYS	Eijsden	60	111	6	48	8	Meuse
HAV	Stad aan 't Haringvliet	53	166	11	67	6.1	Meuse
HEU	Heusden	60	62	6	26	4.3	Meuse
ROO	Roosteren	12	89	9	52	5.8	Meuse
NAM	Nameche	64	40	3	20	6.7	Meuse
TAI	Tailfer	49	39	4	21	5.3	Meuse
STV	Stevensweert	62	114	10	75	7.5	Meuse
KEI	Keizersveer	54	177	13	68	5.2	Meuse
LUI	Luik	63	57	4	17	4.3	Meuse

Each cluster was given an unique identifier, consisting of the name of the location and the number of the cluster. The occurrence of a cluster with specific substances proved, in some cases, unique for a location. Other clusters are recurring in multiple locations, such as the clusters shown in Table 4. These clusters are termed 'Recurring clusters'. Some of the clusters are based on a single measurement in which the substances were suddenly unexpectedly high. These clusters could indicate 'incidents'. Several individual substances that are measured in Rhine and Meuse locations *never* occurred in a cluster, like acetaminophen (paracetamol) and trichloroacetic acid (not shown). This points to erratic emissions for such substances. Other substances were *always* member of a cluster, in any location, such as titanium or indeno(1,2,3-cd)pyreen (not shown). The file with all clusters and the substances can be found in the data package (https://doi.org/10.5281/zenodo.8220952 2023) associated with this report.

These results provide a benchmark for clustering analyses. If a clustering analyses is done in a new location, it can be compared to the results above. Questions can be asked with regard to the percentage substances in clusters (indicative of the structural influences in the location on the substances) or the number of 'incidents' to quantify unexpected short episodes of pollution. Also the specific composition of substances in clusters can be compared.

3.3 Associating extra information to clusters

The fact that substances are present in clusters means they do not vary in concentration at random. Some underlying cause must result in their similar varying concentrations. To find such causes, the clusters are linked to three possibly determining factors: emissions or uses (by finding overlap with Reference lists), environmental conditions at times when concentrations are particularly high (e.g., heavy rainfall, temperature, river discharge, windspeed), and substance properties (e.g. solubility, Kow) (see Figure 5, and Table Appendix IV).

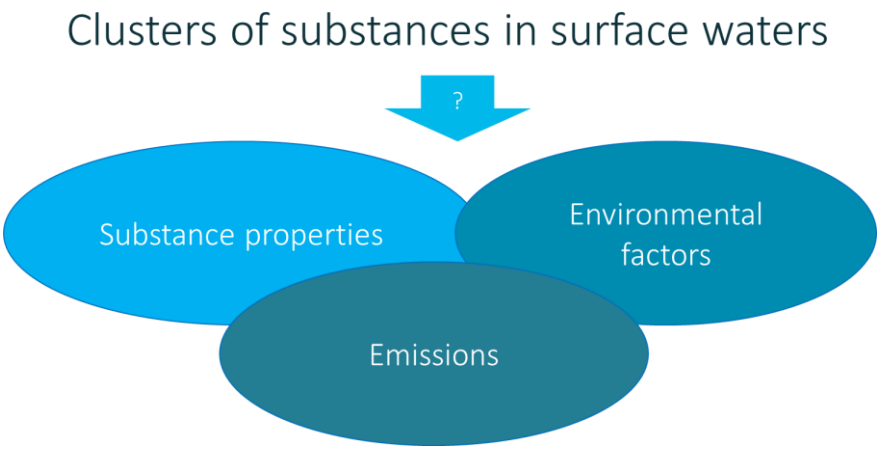


Figure 5. Clusters of substances and information that could lead to the causes of the substances' concerted presence at any time in any place.

3.3.1 Linking substance properties to clusters

Properties of substances, such as summarized in Table Appendix IV, influence the likelihood of finding substances in surface water at any time or place. For instance, the tendency of a substance to migrate from water to air via volatilization will depend on its water solubility, vapor pressure, Henry's law constant and its concentration. In contrast, the tendency of a substance to enter the water system via overland flow depends on Koc and solubility. The tendency of a substance to bind to sediment in water depends on solubility, Kow, Koc and density (*Agency for Toxic Substances and Disease Registry*). This may also influence the tendency of the substance to accumulate on river banks.

For *all* clusters with more than 4 substances, not being organic (because no properties could be retrieved via the prediction softwares), we investigated if the clusters can be linked to the properties (see Table Appendix IV) of the substances in the cluster. The results of these analyses are in the figures in Appendix IV. Figure 6 is shown here as an example. In Figure 6 the log Solubility values of all relevant clusters of all locations are shown. Several clusters are clearly associated with very low log Solubility, compared to the log Solubility that is average for all substances (both in and out of clusters) which is indicated in the green boxplot. Solubility is an important indication of a contaminant's mobility in the aquatic environment, and its ability to reach drinking water sources such as groundwater. A low solubility makes a substance less mobile. From Figure 6 it can therefore be derived that some clusters are specifically associated with the low solubility of the substances in the cluster.

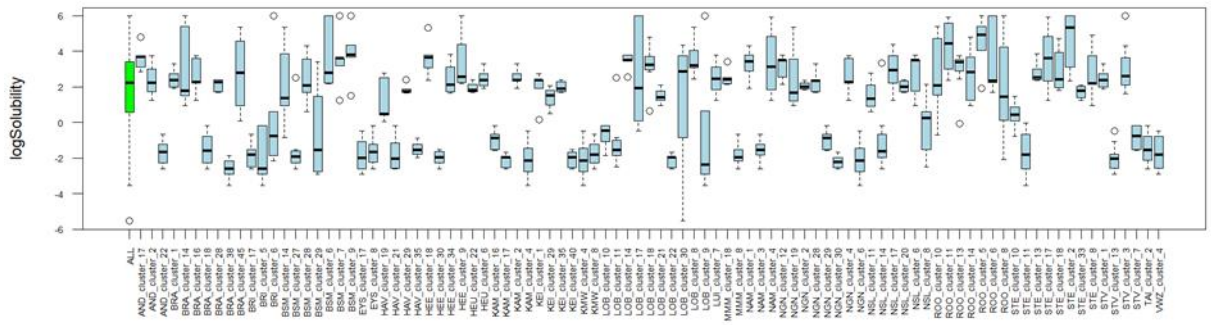


Figure 6. The association of clusters with log Solubility. The green box indicates the average log Solubility value of all the substances, also those that do not appear in a cluster. Every blue box indicates one cluster of substances, named on the x-axis by a location code and a cluster number. See Table 2 for the abbreviations of the cluster location codes.

Other results in the Figures in Appendix IV show that in several clusters consist of substances that:

- have remarkably **low solubility** compared to average values of substances in the dataset
- have remarkably **high Koc** compared to average values of substances in the dataset
- have remarkably **high half-life** compared to average values of substances in the dataset
- have remarkably **high or low volatility**-related properties such as Volatility and Henry's constant.
- have remarkably **high atmospheric hydroxylation constant** (AOH)

This indicates that these are important properties that influence the fate of substances; these may contribute to the similar concentration patterns of substances in some clusters. Other properties (see Appendix IV) may possibly incidentally cause a cluster. The density, for instance, is a property of a chemical in its pure form, and is less relevant when a substance is dissolved in water. It is relevant however for liquids that can float on water because of this property.

In contrast, if a single cluster consist of substances that display a wide range of a property, it can be an indication that the emission is a current and repeating emission. For instance, LOB_cluster_30 has a wide range of solubilities (Figure 6). It consists of two anti-epileptics (primidone and lamotrigine), two high blood pressure regulators that also aid in kidney function with diabetes (telmisartan and candesartan), one antibiotic (Sulfamethoxazole) and an artificial sweetener (sucralose). These could be caused by a regular emission from a sewage treatment system with either patients or a health care center near the monitoring point.

3.3.2 Linking environmental conditions to clusters

For all clusters (again, only those with four or more substances) it was investigated if high concentrations in the clusters can be linked to environmental conditions. The results of these analyses are in the figures in Appendix IV. Figure 7 is shown here as an example. In Figure 7 the river discharge values associated with the highest concentrations of substances per cluster are shown for the river Meuse. Several clusters are clearly associated with very high discharge in the Meuse. Others are associated consistently with low discharge, compared to river discharge that is average for high concentrations of all substances (both in and out of clusters) which is indicated in the green boxplot.

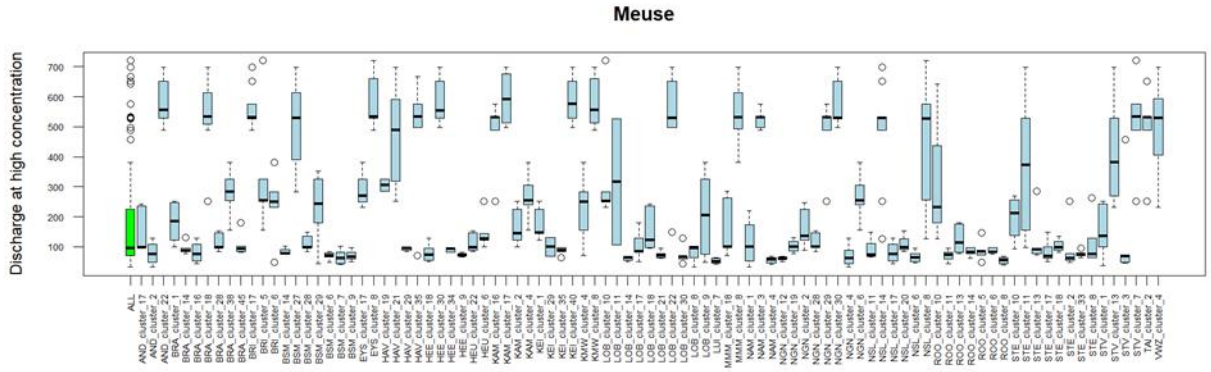


Figure 7. The association clusters with river discharge (m3/s) in the Meuse. The green box indicates the average river discharge value of all substances when their concentration is high (top 10 percent), also those that do not appear in a cluster. All blue boxes indicate one specific cluster, indicated on the x-axis by a location code and a cluster number. See Table 2 for the abbreviations of the cluster location codes.

Box 1. Influence of varying river discharge The clusters are made based on normalized measured substance concentrations. The unit 'concentration' (µg/l) has toxicological relevance, so this makes sense. Another approach to quantify a substance in a water system is the load (µg/s), which is calculated by taking the product of the concentration and the associated river discharge (l/s). A varying discharge typically results in variations in the measured concentrations of substances due to dilution effects. If the load is constant, at a very low discharge a concentration is expected to be higher due to low dilution, and vice versa. In this report, we consider this as simply one of the potential causes for the clustering of substances.

Some naturally occurring substances are dissolved in run-off water and this counteracts their dilution; their concentrations may remain constant with increasing precipitation and associated higher discharge (Ying et al., 2022) (Sjerps et al., 2017). Some substances like chloride can have both natural and anthropogenic origins (Pronk, 2021) with different emission pathways and a mixed relation to discharge. Moreover, other processes can influence the concentration of a substance with discharge. Substances can, for instance, build up when rain is infrequent and have increased run-off at a rain event. This counteracts the effect of dilution in the waterbody only in such situations.

As an example, in Appendix II the Spearman correlation of river discharge with parameters' concentrations can be viewed for an example location. The plot with correlations between parameter loads and discharge indicates that generally loads increase with increasing discharge (indicated by the mostly positive correlations). However, the correlation of concentrations of substances with discharge can vary from positive to negative.

The figures in Appendix IV show that several clusters are associated with one or more environmental conditions:

- Specific clusters were associated with relatively **high or very low temperatures**. The substances in those clusters had high concentrations at high or low temperatures in both Rhine and Meuse. Likewise, these clusters were associated with sun hours and a high or low evaporation potential. This latter condition is dependent on sun irradiation and temperature and this is also dependent on season. This association can be attributed to seasonal use of these chemicals or seasonal differences in degradation / mobilisation of these chemicals leading to seasonal deviations.
- For **river discharge** as a condition, the patterns differed between Rhine and Meuse. For Rhine the association between clusters and river discharge was weak, whereas for the Meuse some clusters have high concentrations at high discharge levels and low discharge levels. The same goes for e-coli and daily precipitation, i.e., the substances in the Meuse are highly affected, while this is less the case for the same substances in the Rhine. The Rhine river carries a lot more water than the Meuse river and effects may be dampened for that reason. The Meuse is rain river whereas the Rhine is also affected by meltwater from glaciers. The dynamics of the Rhine are usually a factor 5 at Lobith (1000-5000 m3/s) while the Meuse has a factor 100 at Eijsden (10-1000 m/s).
- The relation of some clusters to **oxygen level** is quite reproducible between the Meuse and the Rhine. The solubility of oxygen is very temperature dependent. There is more oxygen solved in water in colder conditions. Alternatively, with higher flow speeds in the river systems, more oxygen may be dissolved in the water. This will typically be the case in high river discharge conditions. In addition, oxygen content may drop at high oxygen use for instance by bacteria in summer conditions.

Other environmental conditions did, overall, not associate clearly with clusters according to the analyses:

- For pH anything between 7 and 8 is considered 'normal'. pH in the rivers did not extend beyond such normal values. No clusters had obvious relations with higher or lower pH values.

- The spread in 'normal' dissolved organic carbon (DOC) values associated with high concentrations of substances was large, and no clusters had obvious relations with high or low DOC values.

So, the results indicate that particular conditions may influence the occurrence of substances as clusters. This will be via an interplay with the influence of the properties of the substances in the clusters. Clusters associated with temperature could point to current (non-legacy) emissions.

3.3.3 Linking emissions via Reference lists to clusters

The occurrence of clusters can be partly explained by the overlap with Reference list substances as well. As stated before, the Reference lists hold lists of substances associated with a particular use, emission source, or substance type. Overall, 63 (out of 164) Reference lists (see Appendix I) were significantly overlapping with one or more clusters in the Meuse and Rhine. Table 3 shows the top 5 overlapping Reference lists. Most of the clusters overlapped, as expected for these rivers, with 'waste water processing' -type of reference lists. It is known that the rivers are influences by wastewater. Also 'Dutch rivers' was found often and this is expected because the list is comprised of substances that are structurally found in the Meuse and Rhine. Another Reference list that is often found to overlap is 'Polycyclic aromatic hydrocarbons (PAHs)' and 'Herbicides based on a triazine group'. It appears that both rivers are for the larger part similarly affected by any of the circumstances that result in the presence of these substances (Table 3).

Table 3. Top 5 significantly overlapping Reference lists for clusters from both Meuse and Rhine locations. The Meuse has 11 locations, Rhine has 9 so the Meuse has more clusters in total.

Reference lists	Meuse clusters overlapping	Rhine clusters overlapping
Dutch Rivers	52	45
Waste water treatment plant	54	38
Installations for waste processing or landfills or refinery	23	21
Polycyclic aromatic hydrocarbons (PAHs)	17	15
Industrial chemicals (containing PCBs)	8	10
Herbicides based on a triazine group	12	8

The dendrogram of Reference list similarities In Appendix I can be used to see how alike reference lists are in the substances that they contain. Based on their distance in the dendrogram, reference lists in Table 3 'Industrial chemicals (containing PCBs)' and 'Herbicides based on a triazine group' are, for instance, quite different reference lists from the waste water processing-type of reference lists. This means these may be two separate emission sources or pathways.

Some Reference lists were uniquely (twice or more) overlapping *only* in clusters of the Meuse, such as substances that are members of lists associated with Corn/silage maize cultivation, Seed onions and onion sets, Greenhouse potted plant (Gerbera and Chrysanthemum, Orchids), Fruit and decorative trees culture, Herbicides. For the Rhine these were Herbicides based on anilides, Organochlorine-based insecticides, and Nutrients. Of course, nutrients *are* present in the Meuse, however these apparently do not cluster together to the extent that they do in the Rhine.

Overall, the impression is that clusters in the Meuse are overlapping more often with agricultural type reference lists. Clusters in the Rhine are overlapping more with pharmaceuticals and industry type reference lists. All overlap

between clusters and Reference lists can be found in the data package (https://doi.org/10.5281/zenodo.8220952, 2023).

All in all, linking the reference lists to clusters in Rhine and Meuse for the most part confirmed already known and general influences (wastewater, industrial compound lists, agriculture). These techniques presumably will work better (to be more specific) in smaller, locally influenced waters.

4 Results: Interpretation

4.1.1 Analyzing recurrent clusters in Rhine and Meuse

Recurring (found in multiple locations) clusters in Rhine and Meuse, are of particular interest. Either there is a common emission source of substances in the recurring clusters in all locations, or some other factor (environmental conditions, substance properties) cause the same substances to occur together in several locations. Eight clearly recurring clusters were identified. The clusters in Appendix IV can be linked to the recurring clusters in Table 4 by their cluster ID (e.g. "BRA_28" is one instance of the recurring herbicide cluster). One remarkable feat is that the recurring clusters consist of similar substances like all metals, all pharmaceuticals, all PAHs, etcetera.

Table 4. Recurring clusters in the Meuse and Rhine. Clusters are considered recurring if they contain similar substances in at least 4 locations. Clusters are ranked from very consistent (23 other clusters are similar) to less consistent (4 clusters are similar) (top to bottom). See Table 2 for abbreviations of locations.

Recurring cluster number	Substances (substances in other than the example cluster between parentheses)	Example clusters (number) in locations	Description
RC 1	Aluminium, barium, beryllium, cadmium, cesium, chroom, ijzer, kobalt, koper, kwik, lithium, lood, mangaan, rubidium, thallium,tin, titaan, vanadium, zink, (nikkel, arseen)	MMM_5 AND_6 LOB_20 NGN_18	Metals Sometimes combined with PAH cluster substances
RC 2	Boor, calcium, chloride, kalium, lithium, Magnesium, molybdeen, Natrium, rubidium, strontium, sulfaat, uranium, (bromide, silicaat als Si)	BRI_1 KEI_28 KAM_14 KMW_11	Salts and reactive (alkali) metals
RC 3	benzo(a)antraceen, benzo(a)pyreen, benzo(b)fluorantheen, benzo(ghi)peryleen, benzo(k)fluorantheen, chryseen, dibenzo(a,h)antraceen, fluoranthene, indeno(1,2,3-cd)pyreen, pyreen, (fenantreen, antraceen)	BRI_17 EYS_8 LOB_20	Polycyclic aromatic hydrocarbons (PAHs) (fossil fuel burning) In some clusters together with PCBs
RC 4	Cyanazine, desethyl-terbutylazine, dimethenamide, dimethenamide-p, metolachloor, terbutylazine, (ethofumesaat, metobromuron, linuron)	BRA_28 NGN_28 NSL_20	Herbicides

		HEU_22	
RC 5	2,2',3,4,4',5'-hexachloorbifenyl (PCB 138), 2,2',4,4',5,5'-hexachloorbifenyl (PCB 153), 2,2',4,5,5'-pentachloorbifenyl (PCB 101), 2,2',5,5'-tetrachloorbifenyl (PCB 52), 2,3',4,4',5-pentachloorbifenyl (PCB 118), 2,3,4,5,2',4',5'-heptachloorbifenyl (PCB 180), 2,4,4'-trichloorbifenyl (PCB 28)	NGN_6 KAM_4 KMW_4 BRA_38	Polychlorinated Biphenyls (PCBs) (industrial and commercial applications)
RC 6	1,2-dimethylbenzeen (o-xyleen), 1,2,4-trimethylbenzeen, Benzeen, Ethylbenzeen, methylbenzeen (tolueen), (1,2,3-trimethylbenzeen, 1,3,5-trimethylbenzeen, 2-ethyltolueen, Ethenylbenzeen, n-propylbenzeen)	KAM_2 BRA_1 HEU_6 KEI_1 NGN_2	Aromatic hydrocarbons (petrol oil and fuel)
RC 7	10,11-dihydro-10,11-dihydroxycarbamazepine, carbamazepine, oxazepam, primidone, sulfamethoxazool, temazepam	NGN_4 AND_2 BRA_16 NSL_17	Pharmaceuticals
RC 8	Amidotrizoïnezuur, ethyleendiaminetetra-ethaanzuur (EDTA), jopamidol, jopamidol, joxitalaminezuur (jopromide, johexol)	NGN_19 AND_1 BRA_14	Contrast-agents

For the recurring clusters of Table 4, the substance properties and conditions in Appendix IV provide insight and a possible explanation why substances in clusters have a similar pattern.

- Clusters consisting of PAHs and PCBs (see Table 4) all have low aqueous solubility and very high Koc (see Appendix IV). In other words, the PAH and PCB substances are not mobile and tend to bind to soil and sediment. This means the substances at the stage of their emission are likely not transported for long stretches over land or through the soil, and probably bound to sediment soon after reaching water. A small difference between PAHs and PCBs is in their persistence in the environment. The half-life, a measure for stability and persistence in the environment is high for both PAHs and PCBs, but highest in PAHs according to the model calculations. The atmospheric hydroxylation rate (AOH) is low for PCBs and high for PAHs. Both cluster types are associated with low temperatures and medium to low evaporation potential and high river discharge and daily precipitation (mostly apparent in the Meuse), although PAH are associated with the highest discharges compared to PCBs. Both cluster types are clearly associated with high oxygen content in the river, which is expected to increase with high river discharge or colder temperatures. This points to regular recurrence of these substances due to resuspension of sediment (Friedman et al., 2011; Guigue et al., 2017; Schneider et al., 2007) caused by high water levels in winter. Increasing levels of PAHs and PCBs are also associated with increasing metal concentrations, which is consistent with the sediment being a repository of metals entering aquatic environments. While release of organic matter into the water column may be expected upon sediment resuspension (Guigue et al., 2017; Komada and Reimers, 2001), the general relation between measured dissolved organic carbon (DOC) and high concentrations of any cluster are not very strong (Appendix IV). It is unclear why this is the case.
- Aromatic hydrocarbons in the recurring cluster (see Table 4) have normal range solubility, Koc and half-life. Calculated biodegradation is relatively high. Also, these substances have a high vapor pressure and low octanol–air partition coefficient (KOA). The density is also low, which means that if the substance in these clusters are in a liquid state (e.g. oil) will float on the water. Combined with the high volatility, the substances

will tend to partition to the air. No specific environmental conditions are associated with the occurrence of high concentrations of these clusters, although precipitation is relatively low when measured concentrations of substances in these clusters are high (see Appendix IV).

- Pharmaceuticals and contrast agents in the recurring clusters (see Table 4, RC 7 and RC 8) have properties like solubility and persistence in a broad range (Appendix IV). Volatility is relatively low. Contrast agents have a relatively low Koc and half-life. This means these are mobile, relatively persistent substances. The clusters are most clearly related to environmental conditions of low to medium river discharge and low precipitation (see the figures of Appendix IV). This means that these clusters emerge when water levels are low and probably a lack of dilution increases concentration levels in these recurring clusters. This would indicate that the load (and thus emissions) of the substances in these clusters is generally rather constant. Namely, constant emissions are independent of river discharge and thus will dilute at high discharge and rise with low discharge. That would fit a constant emission, not dependent on season, via treated wastewater (Paíga et al., 2016). This makes sense as many pharmaceuticals are consumed in stable volumes over the year, while for some there are seasonal trends in consumption such as antibiotics and pharmaceuticals related to seasonal infectious diseases (Azuma et al., 2012).
- Herbicides in the recurring cluster (see Table 4) have all substance properties within average values (Appendix IV). Also the river discharge and precipitation when the concentrations of substances in these clusters is high, is in the normal range. The clusters are associated strongly though with high temperatures, evaporation potential, sun hours and low oxygen levels in the river. This points to seasonal reoccurrence in summertime (Gusmaroli et al., 2019; Hladik et al., 2014).

For all the recurring clusters of substances with average/high solubility and average/low Koc the route from point of emission to monitoring may be long, because the substances are relatively mobile and do not tend to bind strongly to sediment. These substances may be caused by 'current' emissions that do not linger but pass by. Recurring clusters containing substances with varying or short half-lives could be current. This is because in a 'legacy' cluster some substances with short half-lives will have disappeared already.

4.1.2 Analyzing a location for apparent clusters of pollution

Individual locations can be analyzed on substances in clusters. These are not necessarily the recurrent clusters (Table 4) but can be specific for a location. As an example we analyze a location in the Rhine catchment: Nieuwegein.

Firstly, the monitoring samples in Nieuwegein seem to cluster according to both season and years (Figure 8). This implies that in addition to effects of season, the pollution pressure has changed between years. Secondly, many clusters were identified as one of the recurring clusters in Table 4. Thirdly, Nieuwegein was characterized by a relatively large number of clusters that may be associated to incidents, these are indicated in red in Figure 8. Some of these clusters – mainly associated with agricultural applications (i.e., pesticide, insecticides, herbicides, and fungicides (clusters 31, 49 and 50) – did not result in a significant overlap with any Reference list. In contrast, cluster 28 significantly overlapped with the lists "Herbicides based on amides" (for controlling weeds in specific crops like potatos) and "Herbicides based on a triazine group" (controlling particular plants' growth by photosynthesis inhibition). This cluster resembles the recurring cluster RC 4 (Table 4), recurring more frequent in locations like Nieuwersluis and Heusden. This cluster has relatively high concentration in summer (Chidya et al., 2022; Pan et al., 2023; Rodríguez-Bolaña et al., 2023).

Cluster 39 significantly overlapped with the reference list "Industrial solvents", and cluster 2 with the lists "Petrol additives", "Industrial solvents", "Motor fuel leakage", and "Industrial chemicals". The latter cluster resembles recurring cluster RC 6 (Table 4). No specific season was associated with the cluster between the different locations. Substances overlapping with the lists "Antidepressants and narcotics", "Domestic wastewater", "Pharmaceuticals", and "Waste water treatment plant" were found in cluster 4, indicating a clear contribution of WWTP at this location (Figure 8) (Osorio et al., 2016; van der Aa et al., 2013). This is expected as the catchment of the Rhine receives wastewater effluents at numerous locations.

Perfluorinated substances such as PFHpA, PFNA, and PFDA (cluster 75) were found to be closely related to cluster 4 (according to the dendrogram) and suggested that these PFAS may be also associated to the same emission source (i.e., WWTP effluents) (Lenka et al., 2021). These two clusters showed very consistent temporal patterns indicating overall higher concentrations in the period 2020 - 2021 than the period 2017 - 2018. A very similar pattern was found also for atenolol, metoprolol, sotalol, furosemide (pharmaceuticals), jopromide (contrast agent) and imidacloprid (insecticide), which are substances typically associated with emissions from WWTP (Wolf et al., 2004), however, these compounds were not identified as a cluster (we named this cluster 01; Figure 8). The emissions were highest in early spring and fall of 2020-2021.

Cluster 19 also included substances that overlapped with the list “Waste water treatment plant” (and “Contrast agents”), however, this cluster appeared to followed a different temporal pattern than clusters 4 and 75, i.e., higher concentrations were mainly observed in winter in 2017, 2018 and 2019 (but also to a lower extent in winter and spring in 2020 and 2021).

Based on visual assessment, a cluster including chloridazon (herbicide), metabenzthiazuron (herbicide), fenazon (anti-inflammatory), aminomethylfosfonzuur (AMPA) (metabolite of glyphosate) and PFAS (PFPeA, PFHxA) was identified (we named this cluster 02; Figure 8). This cluster contained a mixture of different chemicals, however, their temporal trend appeared to be similar to that observed for clusters 4 and 75, which were linked to WWTP emissions. Because wastewater is emitted constantly, concentrations from those emissions are in principle expected to be the highest with low river discharges, that generally occur in summer and autumn. The observed pattern is opposite to what is expected when emissions are constant and concentration dynamics are determined by dilution. Possible reasons for such a deviating pattern can be seasonal use or emissions of these substances, or variable removal in by microbes in the wastewater treatment plants related to temperature. This remains, as of yet, unclear and might even differ between the chemicals within the cluster as long as the temperature and river discharge conditions are closely correlated. The listed PFAS are very persistent, so variable biodegradation is not expected, making the hypothesis of seasonal use or emissions more suitable.

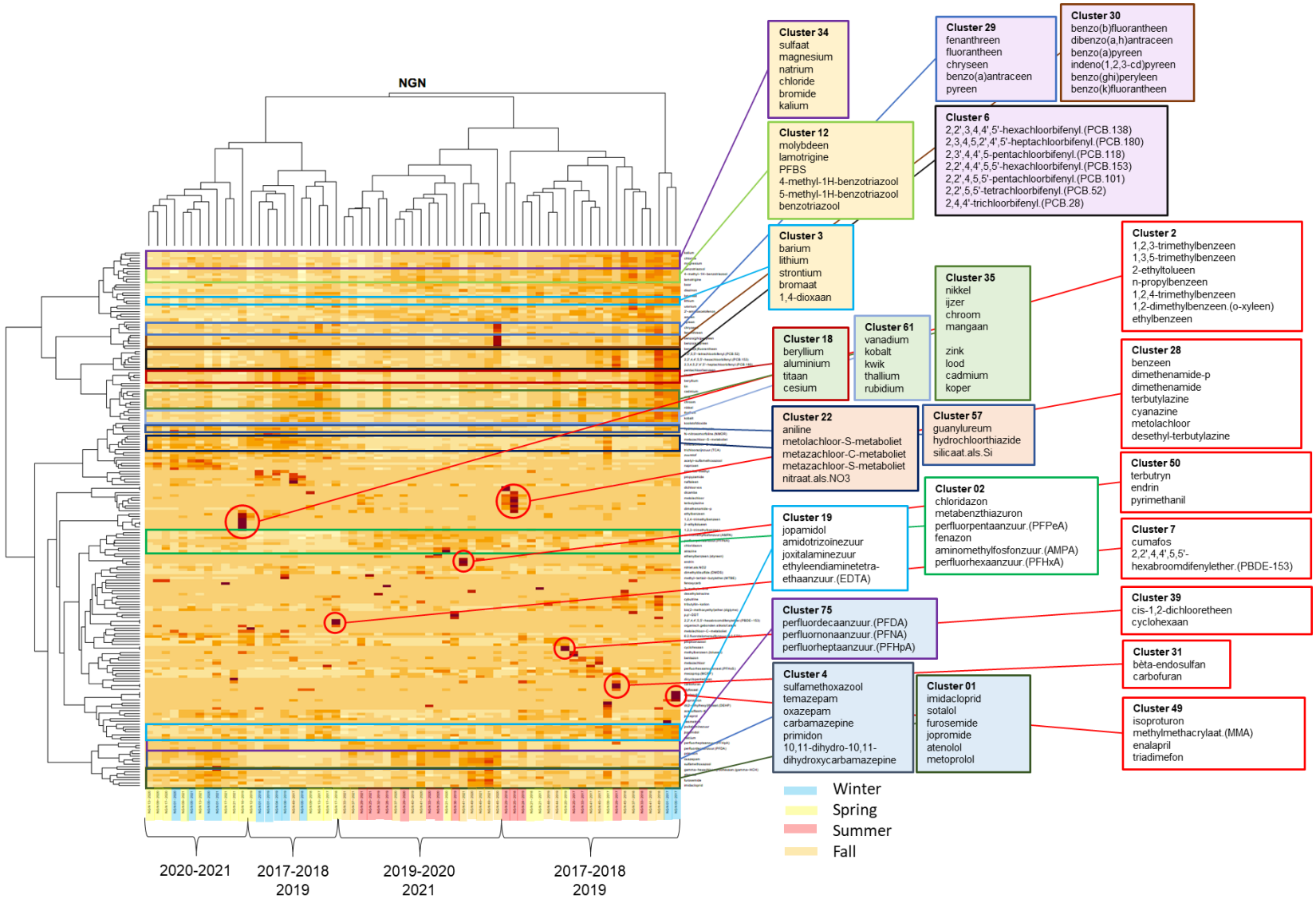


Figure 8. Heatmap obtained using data from Nieuwegein. Boxes with the same background color indicate clusters that may be combined into a larger cluster. Red boxes indicate a single 'incident'.

4.1.3 Case study: individual substance

For individual substances, the context of other substances in their clusters can provide hypotheses towards their source or emission route. In this paragraph we discuss arsenic and benzotriazole, as example substances.

- Arsenic (7440-38-2) is a member of several very general reference lists: Dutch Rivers; Installations for waste processing or landfills or refinery; Waste water treatment plant. Clusters containing arsenic give more detailed information. At two different locations, arsenic is in clusters that are overlapping with the reference list 'Herbicides based on a triazine group' (Nieuwersluis, Brakel). In one location (Lobith) arsenic is associated with a number of Polycyclic aromatic hydrocarbons (PAHs) which is a known common co-contamination (Sun et al., 2018). In another location the arsenic is associated with PCBs. In one location (Stevensweert), the clusters containing arsenic are associated with metals. The properties of the associated substances are different between the clusters. The clusters that are overlapping with herbicide lists consist of highly soluble, medium Koc, low half-life substances. The clusters that are overlapping with lists of PAHs or PCBs are, in contrast, consisting of low soluble substances with high Koc and high half-life. For the cluster with metals, no data on substance properties is available. The differences in substance properties within the clusters that contain arsenic means that there could be different conditions or emissions between the locations that cause a co-occurrence of the substances with arsenic. Arsenic is, for instance, a constituent of potent all-round arsenic herbicide (Qi and Donahoe, 2008; Whitmore et al., 2008). This leads to the hypothesis that use as herbicide may cause the presence of arsenic in the locations where the overlap with Reference lists points that way. The use of arsenic herbicides has been curtailed in most developed countries, though. This means that it may be legacy contaminations. On the other hand, the ionization state of arsenic may be an alternate explanation why it occurs in different cluster types. Fakhreddine et al. (2021) describe how arsenic of geogenic origin can appear in surface water, for instance.

- Benzotriazole (95-14-7) is known for its great versatility. It is used amongst others in antifreezes, heating and cooling systems, hydraulic fluids, vapor-phase inhibitors, as anti-corrosive and drug precursor. The substance is a member of several Reference lists: Treated wastewater; Dutch Rivers; Industrial chemicals (benzotriazoles); Chronomarker (Winter). The substance is in several clusters in Meuse and Rhine. The overlap of the clusters is generally with wastewater related lists (Reemtsma et al., 2010; Weiss et al., 2006), sometimes combined with an extra, more specific reference list such as 'pharmaceuticals' or 'industry' or 'benzotriazoles'. None of the clusters where benzotriazole is member contain substances with remarkable properties (see as an example Figure 9 for the property Koc). Moreover, none of the substances in the clusters have high concentrations at any particular environmental condition, although concentrations tend to be high at low river discharge. This leads to the hypothesis that benzotriazole is detected in the water system because of direct emissions by wastewater treatment plants.

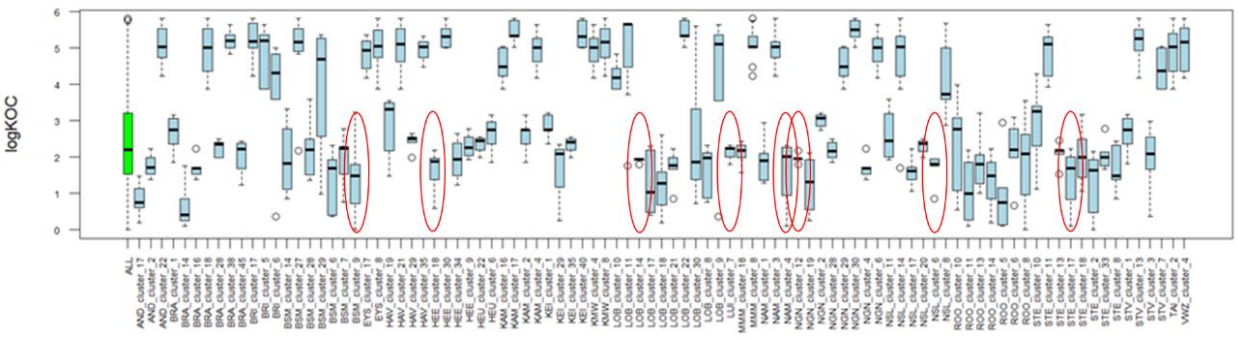


Figure 9. logKOC of substances in clusters. Red circles indicate the clusters that contain benzotriazole. None are remarkably different from the 'average' range of logKOC (green boxplot).

5 Discussion

In this report data tools were used to aid in the knowledge why varying concentrations of substances are found at any location at any time. The data tools are unbiased, they find structures based on data. Environmental scientists can understand, or try to understand the observations, bringing the unbiased approach and the mechanistic understanding based on prior knowledge together.

Firstly, we applied statistical tests to identify clusters in datasets with concentration data. Cluster analysis revealed groups of chemicals that consistently occurred in multiple samples of a location, and these may be indicative of a specific source, origin, fate, or cause of pollution.

Secondly, we determined overlap of substances in the cluster with emissions via Reference lists and we associated properties of substances to the cluster as well as environmental conditions. This aids in the interpretation of the clusters and in formulating hypotheses on whether the clusters were current emissions or legacy emissions, if they were mobile and persistent, etcetera. Overlap of clusters with Reference lists gave an indication of the possible source, origin, use or substance class in the cluster. The Reference lists were based on a review of the available scientific literature, as well as relevant national and EU documentation. This list contains 164 sub-lists of emission sources, substance types, or specific uses to which almost two thousand substances are associated. In the future, other Reference lists can be constructed. For example substances in permits of individual companies. In Appendix VII some other possible Reference lists or indicator substances are listed after a literature review.

Clusters of substances with consistently varying concentration across different samples were found for each of the eighteen locations in the Meuse and Rhine. Many clusters could be statistically linked to a combination of substance properties, environmental conditions, and Reference lists. Some clusters were recurring in several different locations. Especially these recurring clusters were seen under specific circumstances like high temperatures (herbicides), low precipitation (aromatic hydrocarbons), high river discharge, high daily precipitation and high oxygen content (PCBs, PAHs), low to medium river discharge and low precipitation (contrast agents and pharmaceuticals). This information can be used in the more detailed interpretation of the occurrence of individual substances. Of course, the environmental conditions are not independent. For example, at low temperatures more oxygen can be dissolved in water.

In addition to environmental conditions, substances in the recurring clusters had specific properties that were higher or lower than average values. The properties of substances are also not independent. A high solubility, for instance, is known to be negatively related to Koc. This should be taken into account in deriving particular conclusions on the role of properties in cluster formation. Nevertheless, these properties provide a possible explanation to why these substances in recurring clusters remain similar in concentration over the different samples under different circumstances. If these are ‘legacy’ substances, finding back the original emission pathway(s) is of course very difficult. Other clusters had substances with variable properties, these could be ‘current’ and more local emissions. Clusters containing substances with low Koc and high half-life could have been current emissions that traveled further, because the substances are persistent and mobile.

With the clusters, the substance properties, conditions, overlap with Reference lists, and the actual temporal concentration variation over the samples, a hypothesis can be formed for every cluster that is found. For the recurring clusters with PAHs and PCBs, for instance, these factors point to resuspension from sediment. This is pointed out earlier in literature (Echols et al., 2008; Gomes et al., 2013; Zhao et al., 2021). For Herbicides, results point to a seasonal application with transport though air.

A challenge with the data from the Rhine and Meuse is that these waterbodies integrate many sources of pollution. The water flows and clusters of substances that simultaneously entered the water can become separated along the distance that is traveled, by degradation, differences in solubility, volatility, or the tendency to stick to organic matter. Also, if some substances are emitted by an industry at some point and other substances by another industry at another point, it would be hard to trace back what the original emission type was. The number of substances measured in Rhine and Meuse is relatively large and diverse, among other things because of wastewater influences in which many sources are integrated. Reference lists derived from wastewater emissions often overlap. By working with clusters we were able to split the data into groups of coherent substances, and overlap with Reference lists was more specific. This worked remarkably well, considering the manifold influences in these river systems. Even so, it can be expected that local surface waters in smaller catchments, such as contributories of the Meuse and Rhine have clusters that point to more specific sources, uses or substance class clusters than the Rhine and Meuse clusters.

A point of attention in interpreting clusters from the heatmap is that the color of the sample does not indicate the actual concentration. A darker color indicates that the concentration was higher in that sample than in other samples. Incidents may theoretically be of low concentration, as long as the base concentration is even lower. This means that to find out if a dark color is a real 'calamity', a follow up is necessary by checking the actual concentrations of substances in clusters. This could result in the observation that a calamity is merely a relatively high concentration compared to what is normally measured in the location.

For the current clusters that are found in the Meuse and the Rhine, there are several other applications possible.

- Study the frequency and nature of calamities (episodes, regular or single occurrence of unusual high concentrations of combined substances).
- The clusters themselves can be a study object. Do the same clusters occur in different waterbodies? Do clusters change over time?
- Significant overlap of clusters of substances with specific Reference lists can be visualized on a map. This provides spatial information of potential contamination, or on a time-scale providing temporal information. We did not include Reference lists based on permits, or measured emissions by specific industrial companies. This is a possibility, though.
- Causes for high concentrations of individual substances in the Rhine and Meuse can be evaluated based on the clusters they appear in, the conditions and substance properties.
- Identification of substances that are not monitored, but that would be expected to be found based on their association with substances that they cluster together with (i.e., 'guilty by association'). This can complement monitoring programs at particular locations.

A word of caution for the interpretation of individual substances of this report is that, especially for PFAS substances that occur at very low concentrations, the practice of putting measurements below the reporting limit to zero may have introduced errors in the formation of clusters by changing the pattern of varying concentrations too much from actual concentrations. For most parameters though, a measurement below reporting limit can be assumed to be relatively low/near to zero. Also, the data was used as-is. In the Meuse data some substances are known under different names and this could cause substances to seem incomplete (not measured) in some locations. A thorough check in a follow up project will prevent those substances from being removed from the analyses because of their conceived incompleteness.

We recommend further work focusing on potential implementation at drinking water companies. This could be in the shape of a workflow or application for extracting and visualizing information useful for environmental forensic applications, i.e., tracing back the causes of pollution. This could, in a later stage, be combined with environmental fate modeling tools developed to not only explain where a contaminant came from (i.e., its emission source or emission cause), but also how the substance ended up in a given water body (i.e., emission path). The map of contaminated locations can in addition be overlapped with a map of industrial, agricultural and commercial activities

in the Netherlands to find links between contaminated samples and physical sources causing pollution. In such a follow up a definite list of substances that are of high importance can be selected. The clusters, conditions, substance properties and reference lists can be used to find out in detail why each of these substances is present at locations at specific times. It should even be possible to make prediction models. Brunsch et al. (2019, 2018) showed that it is possible to explain concentration variations of substances in water systems.

Future applications including clustering may also include coupling with non-targeted screening (NTS) analysis to perform 'retrospective screening'. The frequent and simultaneous occurrence of unknown (and known) substances may be used to identify clusters linked to pollution sources and prioritize unknowns for further identification. Another future application would be to quantify the correlation between substance concentrations and environmental conditions and predict the fate of the substance. Or, the concentrations of a substance of interest that was not measured could perhaps be predicted from the concentrations of substances that are clustered with the substance of interest.

Another future application is groundwater. Groundwater is less affected by chemical pollutants than surface water, and is rather stagnant compared to surface water. Therefore, overlap with reference lists will assumably be more specific for each location. It is known that groundwater pollution with substances is not random. McMahon et al. (2022) for instance was able to predicted PFAS concentrations in groundwater using conditions and the concentrations of other parameters. That is why it seems feasible to also look for clusters of substances in groundwater.

6 Conclusions

Monitoring data contain far more information than simply concentration levels that are used for assessing compliance with water quality guidelines. Clustering and the cooccurrence of certain types of chemicals and differences and similarities between locations provide a wealth of information for building and testing hypothesis on sources, emissions and impact of conditions on concentrations and loads. This helps us to formulate new hypothesis and thereby establish better knowledge. It provides an important piece in the iterative puzzle towards understanding sources and their contributions.

In this investigation we focused on monitoring datasets for surface water. We have reviewed, tested and identified statistical tools that are potentially useful for performing environmental forensic investigations that aim at extracting valuable (and often overlooked) information from existing datasets. We have applied these tools to monitoring datasets from surface water and assessed their suitability for the different types of datasets.

Based on the results obtained in this preliminary investigation, statistical analysis and clustering appeared to be useful for processing existing datasets and extracting information that would otherwise remain concealed within their datasets. Although currently only the first steps are taken, this report shows that current monitoring data and applied techniques already provide several leads for ultimately understanding and possibly predicting concentrations of substances. Thereby results can, in the future, support the formulation and evaluation of mitigation strategies.

7 References

Arip, M.N.M., Heng, L.Y., Ahmad, M., Ujang, S., 2013. A cell-based potentiometric biosensor using the fungus Lentinus sajor-caju for permethrin determination in treated wood. Talanta 116, 776–781. https://doi.org/10.1016/j.talanta.2013.07.065

Azuma, T., Nakada, N., Yamashita, N., Tanaka, H., 2012. Synchronous Dynamics of Observed and Predicted Values of Anti-influenza drugs in Environmental Waters during a Seasonal Influenza Outbreak. Environ. Sci. Technol. 46, 12873–12881. https://doi.org/10.1021/es303203c

Baragaño, D., Ratié, G., Sierra, C., Chrastný, V., Komárek, M., Gallego, J.R., 2022. Multiple pollution sources unravelled by environmental forensics techniques and multivariate statistics. J. Hazard. Mater. 424, 127413. https://doi.org/10.1016/j.jhazmat.2021.127413

Brunsch, A.F., Langenhoff, A.A.M., Rijnaarts, H.H.M., Ahring, A., ter Laak, T.L., 2019. In situ removal of four organic micropollutants in a small river determined by monitoring and modelling. Environ. Pollut. 252, 758–766. https://doi.org/10.1016/j.envpol.2019.05.150

Brunsch, A.F., ter Laak, T.L., Rijnaarts, H., Christoffels, E., 2018. Pharmaceutical concentration variability at sewage treatment plant outlets dominated by hydrology and other factors. Environ. Pollut. 235, 615–624. https://doi.org/10.1016/j.envpol.2017.12.116

Buttiglieri, G., Peschka, M., Frömel, T., Müller, J., Malpei, F., Seel, P., Knepper, T.P., 2009. Environmental occurrence and degradation of the herbicide n-chloridazon. Water Res. 43, 2865–2873. https://doi.org/10.1016/j.watres.2009.03.035

Byer, J.D., Struger, J., Sverko, E., Klawunn, P., Todd, A., 2011. Spatial and seasonal variations in atrazine and metolachlor surface water concentrations in Ontario (Canada) using ELISA. Chemosphere 82, 1155–1160. https://doi.org/10.1016/j.chemosphere.2010.12.054

Carabajal, M., Teglia, C.M., Maine, M.A., Goicoechea, H.C., 2021. Multivariate optimization of a dispersive liquid-liquid microextraction method for the determination of six antiparasite drugs in kennel effluent waters by using second-order chromatographic data. Talanta 224, 121929. https://doi.org/10.1016/j.talanta.2020.121929

Chidya, R., Derbalah, A., Abdel-Dayem, S., Kaonga, C., Sakugawa, H., 2022. Ecotoxicological and human health risk assessment of selected pesticides in Kurose River, Higashi-Hiroshima City (Japan). Water Environ. Res. 94. https://doi.org/10.1002/wer.1676

Echols, K.R., Brumbaugh, W.G., Orazio, C.E., May, T.W., Poulton, B.C., Peterman, P.H., 2008. Distribution of Pesticides, PAHs, PCBs, and Bioavailable Metals in Depositional Sediments of the Lower Missouri River, USA. Arch. Environ. Contam. Toxicol. 55, 161–172. https://doi.org/10.1007/s00244-007-9123-0

Fakhreddine, S., Prommer, H., Scanlon, B.R., Ying, S.C., Nicot, J.-P., 2021. Mobilization of Arsenic and Other Naturally Occurring Contaminants during Managed Aquifer Recharge: A Critical Review. Environ. Sci. Technol. 55, 2208–2223. https://doi.org/10.1021/acs.est.0c07492

Friedman, C.L., Lohmann, R., Burgess, R.M., Perron, M.M., Cantwell, M.G., 2011. Resuspension of polychlorinated biphenyl-contaminated field sediment: release to the water column and determination of site-specific KDOC. Environ. Toxicol. Chem. 30, 377–384. https://doi.org/10.1002/etc.408

Gomes, H.I., Dias-Ferreira, C., Ribeiro, A.B., 2013. Overview of in situ and ex situ remediation technologies for PCB-contaminated soils and sediments and obstacles for full-scale application. Sci. Total Environ. 445–446, 237–260. https://doi.org/10.1016/j.scitotenv.2012.11.098

Guigue, C., Tedetti, M., Dang, D.H., Mullot, J.-U., Garnier, C., Goutx, M., 2017. Remobilization of polycyclic aromatic hydrocarbons and organic matter in seawater during sediment resuspension experiments from a polluted coastal environment: Insights from Toulon Bay (France). Environ. Pollut. 229, 627–638. https://doi.org/10.1016/j.envpol.2017.06.090

Gusmaroli, L., Buttiglieri, G., Petrovic, M., 2019. The EU watch list compounds in the Ebro delta region: Assessment of sources, river transport, and seasonal variations. Environ. Pollut. 253, 606–615. https://doi.org/10.1016/j.envpol.2019.07.052

Harman, C., Reid, M., Thomas, K.V., 2011. In Situ Calibration of a Passive Sampling Device for Selected Illicit Drugs and Their Metabolites in Wastewater, And Subsequent Year-Long Assessment of Community Drug Usage. Environ. Sci. Technol. 45, 5676–5682. https://doi.org/10.1021/es201124j

Hermsen, S.A.B., Pronk, T.E., van den Brandhof, E.-J., van der Ven, L.T.M., Piersma, A.H., 2013. Transcriptomic analysis in the developing zebrafish embryo after compound exposure: Individual gene expression and pathway regulation. Toxicol. Appl. Pharmacol. 272, 161–171. https://doi.org/10.1016/j.taap.2013.05.037

Hillebrand, O., Nödler, K., Licha, T., Sauter, M., Geyer, T., 2012a. Caffeine as an indicator for the quantification of untreated wastewater in karst systems. Water Res. 46, 395–402. https://doi.org/10.1016/j.watres.2011.11.003

Hillebrand, O., Nödler, K., Licha, T., Sauter, M., Geyer, T., 2012b. Identification of the attenuation potential of a karst aquifer by an artificial dualtracer experiment with caffeine. Water Res. 46, 5381–5388. https://doi.org/10.1016/j.watres.2012.07.032

Hladik, M.L., Kolpin, D.W., Kuivila, K.M., 2014. Widespread occurrence of neonicotinoid insecticides in streams in a high corn and soybean producing region, USA. Environ. Pollut. 193, 189–196. https://doi.org/10.1016/j.envpol.2014.06.033

Kahl, S., Nivala, J., van Afferden, M., Müller, R.A., Reemtsma, T., 2017. Effect of design and operational conditions on the performance of subsurface flow treatment wetlands: Emerging organic contaminants as indicators. Water Res. 125, 490–500. https://doi.org/10.1016/j.watres.2017.09.004

Kasprzyk-Hordern, B., Baker, D.R., 2012a. Estimation of community-wide drugs use via stereoselective profiling of sewage. Sci. Total Environ. 423, 142–150. https://doi.org/10.1016/j.scitotenv.2012.02.019

Kasprzyk-Hordern, B., Baker, D.R., 2012b. Enantiomeric Profiling of Chiral Drugs in Wastewater and Receiving Waters. Environ. Sci. Technol. 46, 1681–1691. https://doi.org/10.1021/es203113y

Kimes, P.K., Liu, Y., Neil Hayes, D., Marron, J.S., 2017. Statistical significance for hierarchical clustering. Biometrics 73, 811–821. https://doi.org/10.1111/biom.12647

Komada, T., Reimers, C.E., 2001. Resuspension-induced partitioning of organic carbon between solid and solution phases from a river–ocean transition. Mar. Chem. 76, 155–174. https://doi.org/10.1016/S0304-4203(01)00055-X

Lenka, S.P., Kah, M., Padhye, L.P., 2021. A review of the occurrence, transformation, and removal of poly- and perfluoroalkyl substances (PFAS) in wastewater treatment plants. Water Res. 199, 117187. https://doi.org/10.1016/j.watres.2021.117187

Loraine, G.A., Pettigrove, M.E., 2006. Seasonal Variations in Concentrations of Pharmaceuticals and Personal Care Products in Drinking Water and Reclaimed Wastewater in Southern California. Environ. Sci. Technol. 40, 687–695. https://doi.org/10.1021/es051380x

McMahon, P.B., Tokranov, A.K., Bexfield, L.M., Lindsey, B.D., Johnson, T.D., Lombard, M.A., Watson, E., 2022. Perfluoroalkyl and Polyfluoroalkyl Substances in Groundwater Used as a Source of Drinking Water in the Eastern United States. Environ. Sci. Technol. 56, 2279–2288. https://doi.org/10.1021/acs.est.1c04795

Osorio, V., Sanchís, J., Abad, J.L., Ginebreda, A., Farré, M., Pérez, S., Barceló, D., 2016. Investigating the formation and toxicity of nitrogen transformation products of diclofenac and sulfamethoxazole in wastewater treatment plants. J. Hazard. Mater. 309, 157–164. https://doi.org/10.1016/j.jhazmat.2016.02.013

Paíga, P., Santos, L.H.M.L.M., Ramos, S., Jorge, S., Silva, J.G., Delerue-Matos, C., 2016. Presence of pharmaceuticals in the Lis river (Portugal): Sources, fate and seasonal variation. Sci. Total Environ. 573, 164–177. https://doi.org/10.1016/j.scitotenv.2016.08.089

Pan, X., Xu, L., He, Z., Wan, Y., 2023. Occurrence, fate, seasonal variability, and risk assessment of twelve triazine herbicides and eight related derivatives in source, treated, and tap water of Wuhan, Central China. Chemosphere 322, 138158. https://doi.org/10.1016/j.chemosphere.2023.138158

Pascual-Aguilar, J., Andreu, V., Picó, Y., 2013. An environmental forensic procedure to analyse anthropogenic pressures of urban origin on surface water of protected coastal agro-environmental wetlands (L'Albufera de Valencia Natural Park, Spain). J. Hazard. Mater. 263, 214–223. https://doi.org/10.1016/j.jhazmat.2013.07.052

Qi, Y., Donahoe, R.J., 2008. The environmental fate of arsenic in surface soil contaminated by historical herbicide application. Sci. Total Environ. 405, 246–254. https://doi.org/10.1016/j.scitotenv.2008.06.043

Rakonjac, N., van der Zee, S.E.A.T.M., Wipfler, L., Roex, E., Kros, H., 2022. Emission estimation and prioritization of veterinary pharmaceuticals in manure slurries applied to soil. Sci. Total Environ. 815, 152938. https://doi.org/10.1016/j.scitotenv.2022.152938

Reemtsma, T., Miehe, U., Duennbier, U., Jekel, M., 2010. Polar pollutants in municipal wastewater and the water cycle: Occurrence and removal of benzotriazoles. Water Res., Emerging Contaminants in water: Occurrence, fate, removal and assessment in the water cycle (from wastewater to drinking water) 44, 596–604. https://doi.org/10.1016/j.watres.2009.07.016

Rodríguez-Bolaña, C., Pérez-Parada, A., Tesitore, G., Goyenola, G., Kröger, A., Pacheco, M., Gérez, N., Berton, A., Zinola, G., Gil, G., Mangarelli, A., Pequeño, F., Besil, N., Niell, S., Heinzen, H., Teixeira de Mello, F., 2023. Multicompartmental monitoring of legacy and currently used pesticides in a subtropical lake used as a drinking water source (Laguna del Cisne, Uruguay). Sci. Total Environ. 874, 162310. https://doi.org/10.1016/j.scitotenv.2023.162310

Schneider, A.R., Porter, E.T., Baker, J.E., 2007. Polychlorinated Biphenyl Release from Resuspended Hudson River Sediment. Environ. Sci. Technol. 41, 1097–1103. https://doi.org/10.1021/es0607584

Seitz, W., Winzenbacher, R., 2017. A survey on trace organic chemicals in a German water protection area and the proposal of relevant indicators for anthropogenic influences. Environ. Monit. Assess. 189, 244. https://doi.org/10.1007/s10661-017-5953-z

Sun, L., Zhu, G., Liao, X., 2018. Enhanced arsenic uptake and polycyclic aromatic hydrocarbon (PAH)-dissipation using Pteris vittata L. and a PAH-degrading bacterium. Sci. Total Environ. 624, 683–690. https://doi.org/10.1016/j.scitotenv.2017.12.169

Suzuki, R., Shimodaira, H., 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22, 1540–1542. https://doi.org/10.1093/bioinformatics/btl117

ter Laak, T.L., Kooij, P.J.F., Tolkamp, H., Hofman, J., 2014. Different compositions of pharmaceuticals in Dutch and Belgian rivers explained by consumption patterns and treatment efficiency. Environ. Sci. Pollut. Res. Int. 21, 12843–12855. https://doi.org/10.1007/s11356-014-3233-9

van der Aa, M., Bijlsma, L., Emke, E., Dijkman, E., van Nuijs, A.L.N., van de Ven, B., Hernández, F., Versteegh, A., de Voogt, P., 2013. Risk assessment for drugs of abuse in the Dutch watercycle. Water Res. 47, 1848–1857. https://doi.org/10.1016/j.watres.2013.01.013

Warner, W., Licha, T., Nödler, K., 2019. Qualitative and quantitative use of micropollutants as source and process indicators. A review. Sci. Total Environ. 686, 75–89. https://doi.org/10.1016/j.scitotenv.2019.05.385

Warner, W., Ruppert, H., Licha, T., 2016. Application of PAH concentration profiles in lake sediments as indicators for smelting activity. Sci. Total Environ. 563–564, 587–592. https://doi.org/10.1016/j.scitotenv.2016.04.103

Weiss, S., Jakobs, J., Reemtsma, T., 2006. Discharge of Three Benzotriazole Corrosion Inhibitors with Municipal Wastewater and Improvements by Membrane Bioreactor Treatment and Ozonation. Environ. Sci. Technol. 40, 7193–7199. https://doi.org/10.1021/es061434i

Whitmore, T.J., Riedinger-Whitmore, M.A., Smoak, J.M., Kolasa, K.V., Goddard, E.A., Bindler, R., 2008. Arsenic contamination of lake sediments in Florida: evidence of herbicide mobility from watershed soils. J. Paleolimnol. 40, 869–884. https://doi.org/10.1007/s10933-008-9204-8

Wolf, L., Held, I., Eiswirth†, M., Hötzl, H., 2004. Impact of Leaky Sewers on Groundwater Quality. Acta Hydrochim. Hydrobiol. 32, 361–373. https://doi.org/10.1002/aheh.200400538

Zhao, C., Xu, J., Shang, D., Zhang, Y., Zhang, J., Xie, H., Kong, Q., Wang, Q., 2021. Application of constructed wetlands in the PAH remediation of surface water: A review. Sci. Total Environ. 780, 146605. https://doi.org/10.1016/j.scitotenv.2021.146605

I Reference lists

In the original collected Reference substance lists, a total of 232 separate lists are present. In these lists a total of 1968 unique substances are included. Some of the lists overlap to an extent. This poses a practical problem because a cluster of substances may overlap with several reference lists that are in essence very similar. To avoid this, overlapping lists were merged. Two lists were considered overlapping if the sum of the percentage overlap between the two lists was greater than 130%. For instance, a case where there was 40% overlap of list1 with list2 and 100% overlap of list2 with list1 would be calculated as 140% overlap. A case where list1 overlaps 70% with list2 and list2 overlaps 60% with list1 is calculated as 130% overlap. All reference lists were checked for overlap against all other reference lists. Overlapping lists were merged and renamed by making a combination of the original list names. This resulted in a reduction to 164 separate reference substance lists. The overall similarity of the new reference lists (expressed as the % remaining overlap) is visualised as a hierarchical clustering in the figure below. Some lists are still more related than others.

Table Appendix I A. Sources and sizes of the reference substances lists, before merging overlapping lists.

List source ID	List source name	Source	# sub-lists	# sub-stances
L1	Chemicals used in agriculture types	CBS	58	1715
L2	Chemicals measured near agriculture	LM GBM (data obtained from Deltares)	8	63
L3	Sewage treatment plants (STP)	Watson database (data driven, substances found >0.1 ug/l in >25 STP effluents)	1	83
L4	Trans-border Meuse	RIWA database (data driven, substances in samples of location Eijsden on average >0.1 ug/l)	2	47
L5	Trans-border Rhine	RIWA database (data driven, substances in samples of location Lobith on average >0.1 ug/l)	2	71
L6	Biocides per product type	ECHA database	20	656
L7	Distinguished groups (diverse)	RIWA-Rijn	89	1714
L8	Micropollutants as source and process indicators	Warner et al. (2019)	17	71
L9	EU emissions by industries	EEA Industries Reporting Database	20	128
L10	Veterinary pharmaceuticals in manure slurries	Rakonjac et al. (2022)	2	28
L11	Sources of PFAS in Dutch surface water	Rijkswaterstaat (2020)	11	13
L12	Typical substances in untreated wastewater	Watson database (data driven, substances found abundantly (>25 STP, at least 0.1 ug/l) in influent, not in effluent, and are well removed (>80%))	1	9

L13	Drug waste constituents	RIVM report (2022)	1	62
L14	A list of chemicals in fertilizers	CompTox lists	1	22
L15	Motor fuel leakage chemicals	CompTox lists	1	27
L16	Natural toxins	CompTox lists	1	90
L17	Veterinary drugs	CompTox lists	1	124
L18	Cyanoginosins (from cyanobacteria)	CompTox lists	1	7

Some substances may be highly biodegradable, volatile, or adhere strongly to soil, and thus, less suitable for the scope of this study, because these substances would be less relevant in reference lists for determining the overlap with monitoring data. We checked if there was a visible threshold were substances with high biodegradation (predictions in Biowin3 combined with predictions if a substance was 'Readily Biodegradable'), high logK_{OC} (tend to be immobile), high volatility or Henry's constant (tend to volatilize to air) are not likely found in Rhine or Meuse water. As a source of these chemical properties we used the open source software 'EpiSuite'. We could not find such a threshold that would indicate that substances are always less relevant (are seldom detected in Meuse or Rhine) because of any of these properties (not shown). For this reason all substances were retained in the reference lists.

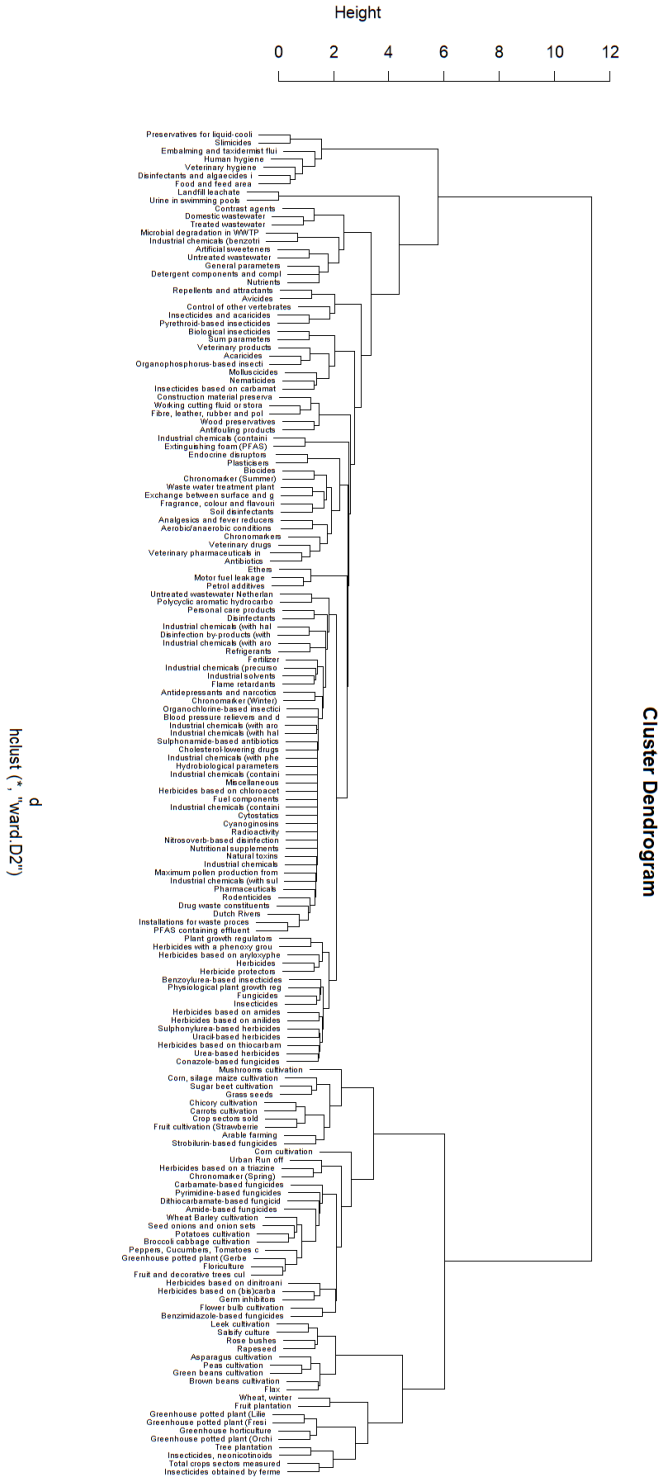


Figure Appendix I A. Reference list hierarchical clustering of percentages overlap. The height of the line indicates the dissimilarity between clusters. The sum of the percentual overlaps between any two lists does not exceed 130%.

II Example of spearman correlations of parameters with river discharge

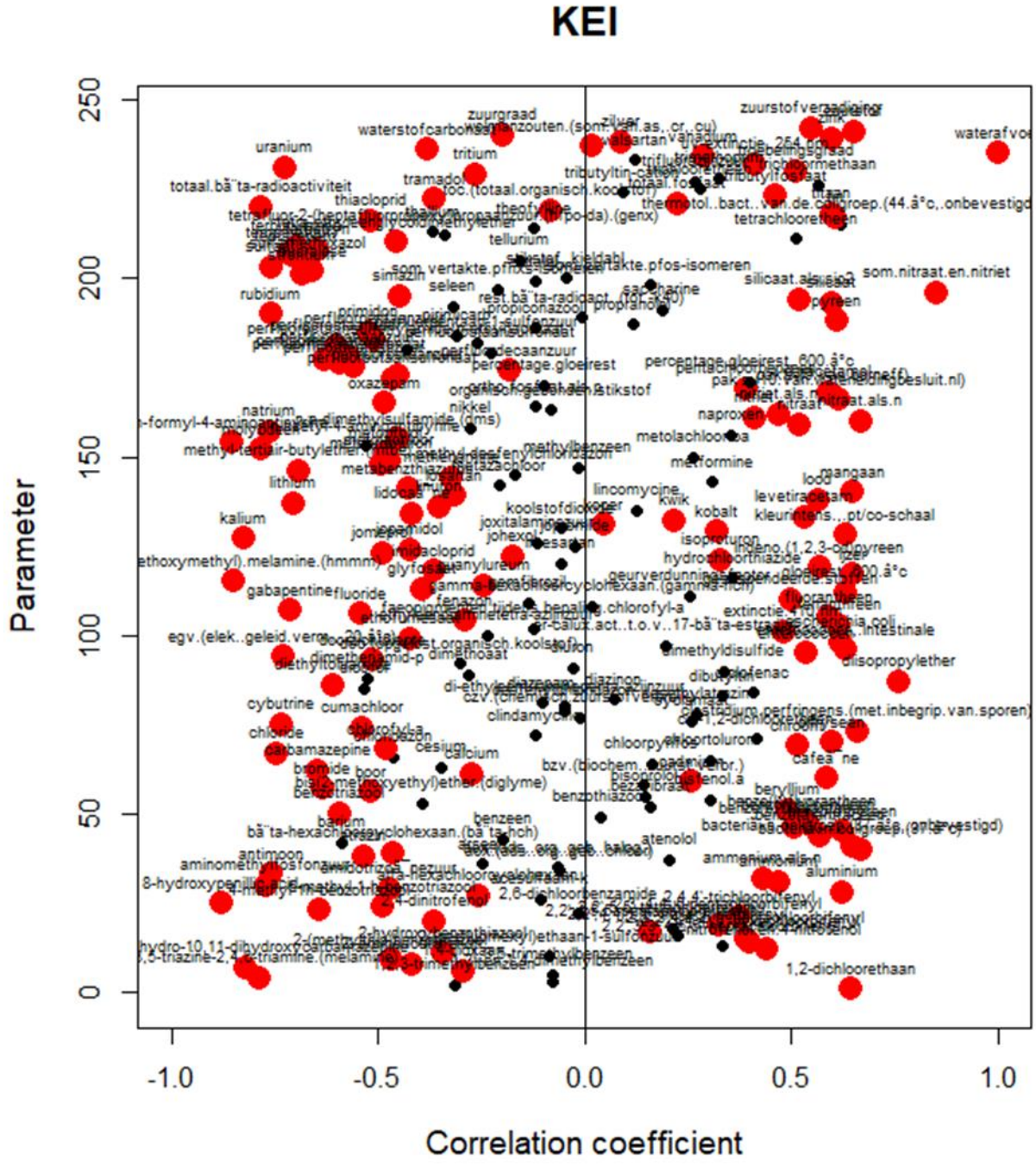


Figure Appendix II A. Correlations of substance concentrations with river discharge ('waterafvoer', upper right corner) for location Keizersveer. Significant correlations are indicated in large red circles. With more observations, lower correlations can become significant. Take-home message of this Figure is the wide range of correlations (negative to positive) of different substances.

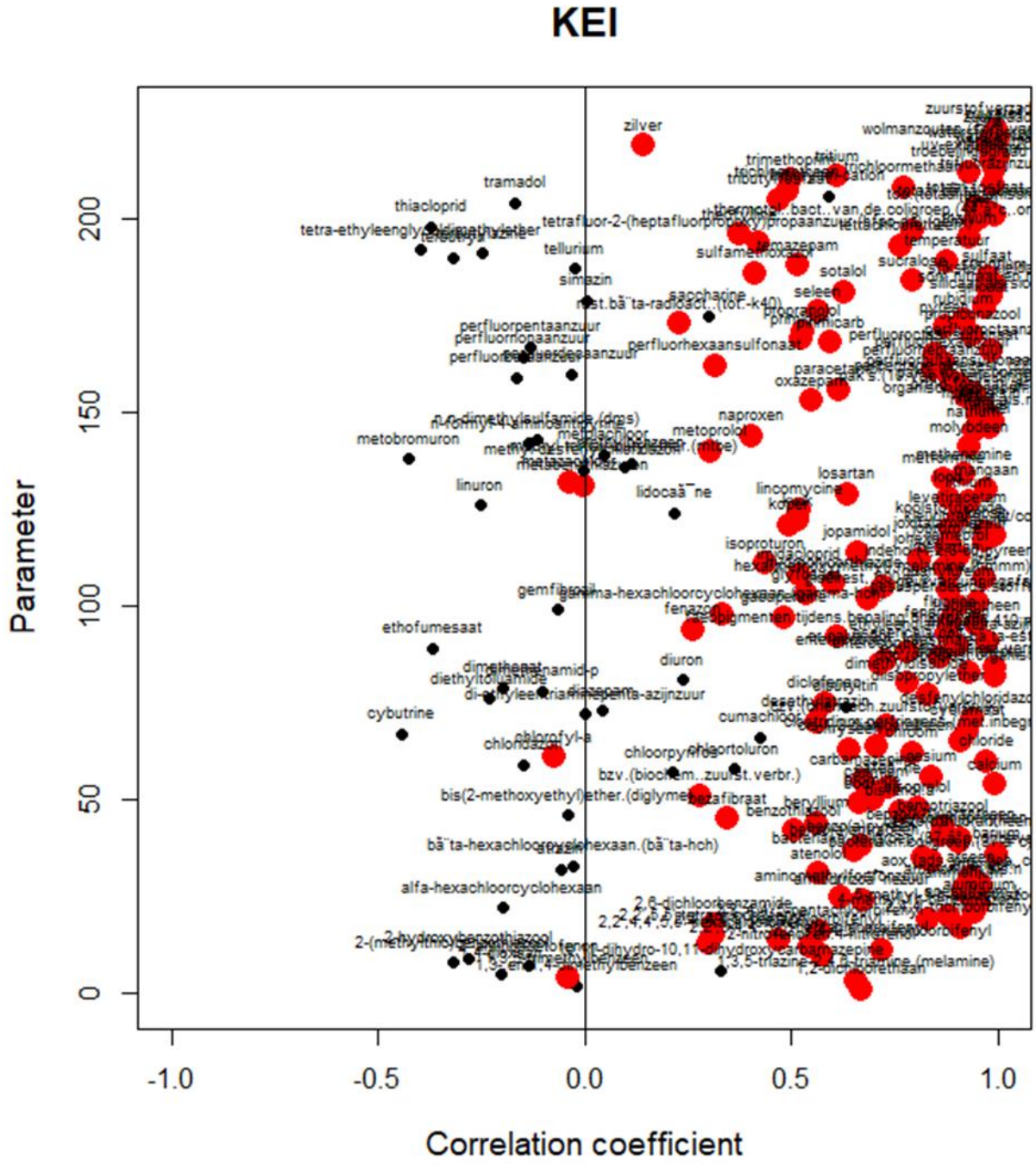


*Figure Appendix II B. Correlations of substances load (per second) with river discharge ('waterafvoer') of location Keizersveer. Load was calculated as: concentration (ug/L) * discharge (L/s). Significant correlations are indicated in large red circles. With more observations, lower correlations can become significant. Take-home message in this Figure is the overall positive correlation between discharge and load (more discharge, more substance). This doesn't necessarily mean that the concentration of the substance is also positively related (see Figure above). So, dilution effects on the concentration can for some substances have higher influence than the increase in load.*

III Comparison of cluster significance methods

Detailed explanation of the ClusSig method

When a hierarchical clustering is computed, it is a given that all substances or samples are in a cluster at any level of cluster numbers. This can be visualised in a dendrogram (see Figure 2 in the main text). At the top, there is only one cluster. One level below, there are two clusters, and so on. At the lowest level, each substance or sample is in its own cluster. For our analyses, the substance clusters are most relevant. Even if all substances are in a cluster at any level, visually some clusters look more consistent than others. Such clusters have substances that are in one cluster towards the bottom of the hierarchy in the dendrogram (Figure 2 in the main text). The clustered normalized concentrations of the substances over the samples can also be visualised in a heatmap. This confirms that some substances are in a cluster with a very similar and consistent pattern over the different samples (e.g. see Figure 8 in the main text).

For our purpose to select clusters of substances that change in concentration together, we want to only select clusters that look consistent in the heatmap and are consistent throughout the dendrogram towards the bottom.(i.e. the point where all substances are in their private cluster). To determine such clusters a 'BTO cluster significance' (ClusSig) method was developed. The steps in the ClusSig method are illustrated in the figure below this text.

The procedure in the ClusSig method is as follows. For each level in number of clusters we randomly (as many times as there are substances) produce a number that represents a cluster. This produces different sized clusters. For example, at a level of 4 clusters with 120 substances we draw the numbers 1-4 120 times. To ensure there is never an empty cluster (as in a real HCA) we initialize the draw with the numbers 1-4 and randomly draw the remaining 116 numbers between 1-4. A result could be that 10 times '1' was drawn, 35 times '2', 45 times '3' and 30 times '4' (total 120). These are the randomly drawn cluster sizes for the 120 substances. For each level of cluster numbers we repeat this 1000 times. A distribution of cluster sizes emerges for each level. Some cluster sizes emerge very frequent (these are logically the average cluster-size for that level), some are rare (very small or very big). We express the distribution of sizes for each level as a 'quantile'. A cluster size at the 90^{th} quantile means that only 10% of all randomly drawn clusters have a bigger size. This means this size occurs not very often in randomly sized clusters. Then, we compare the actual cluster sizes at a level in the HCA with monitoring data with that of the calculated quantiles. Every substance in a cluster at the different levels gets assigned that quantile. Clusters at any level with a quantile-size >90 are considered 'significant'. This quantile level of 90 was selected by comparing the clusters that could be identified visually and via the ClusSig method. One difficulty remains, and that is to determine the optimal cluster number level at which to regard the 'significance' of the clusters. We argue that substances that remain in a cluster at lower levels in the hierarchy are very consistently clustered. At the same time, good sized useable clusters will occur at a level at which many substances are in a high quantile cluster. This is determined by the sum of quantiles at each level. These two arguments lead to the selection of an 'optimal level' where the sum of quantiles start to decline towards the bottom of the hierarchy. This is a 'bending point'. All clusters that are significant at the level of the bending point, or become significant at any level below, are considered significant clusters.

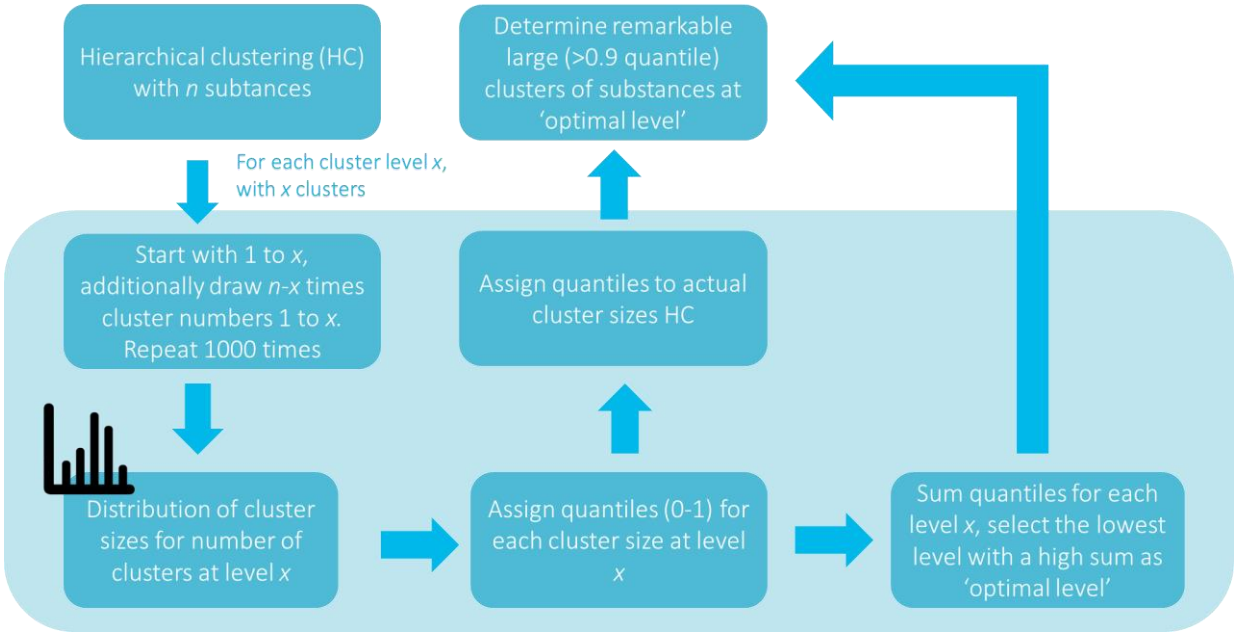


Figure Appendix III A. Flow diagram for determining significant clusters in the ClusSig method.

In short, the ClusSig method works with the assumption that any large cluster compared to an expected size is extraordinary and significant.

This method is a little less sophisticated than the methods in Kimes et al. (2017) and Suzuki and Shimodaira (2006). Kimes et al. (2017) basically test at every junction of the dendrogram if the values of elements in the cluster follow a single Gaussian distribution stronger than a random simulated cluster of that size with an imposed Gaussian distribution, and deciding if that indicates a single cluster. Suzuki and Shimodaira (2006) use a bootstrap method to make many instances of the hierarchical cluster under investigation, and seeing how many times a cluster appears from random sampled elements. If it appears often, it is a robust cluster. So, both use the actual calculated values of elements by the clustering methods in the hierarchy whereas the ClusSig uses only expected size distributions. The use of the ClusSig method instead of established methods is preferred because of the simplicity of the approach (it is understandable) and the flexibility to test and adjust it.

We applied the two other methods for cluster significance that are available in the statistical language 'R', Pvclust (Suzuki and Shimodaira, 2006) and Sigclust2 (Kimes et al., 2017). Pvclust tends to assign significance to small clusters in the data. This is not practical. Sigclust2 assigns significance to both the larger and smaller dense clusters. Unfortunately, the predefined functions in Sigclust2 only allow to determine clusters at a level in the dendrogram were all substances were in significant clusters. These technical limitations made the use of Sigclust2 unpractical even though a very nice visualization was possible and significance seemed accurate. Appendix III provides a comparison of significant clusters between the three methods and a clustering on visual inspection. ClusSig fortunately generally performed as well as the two methods, compared to the clusters that were assigned based on the visual inspection of the heatmap.

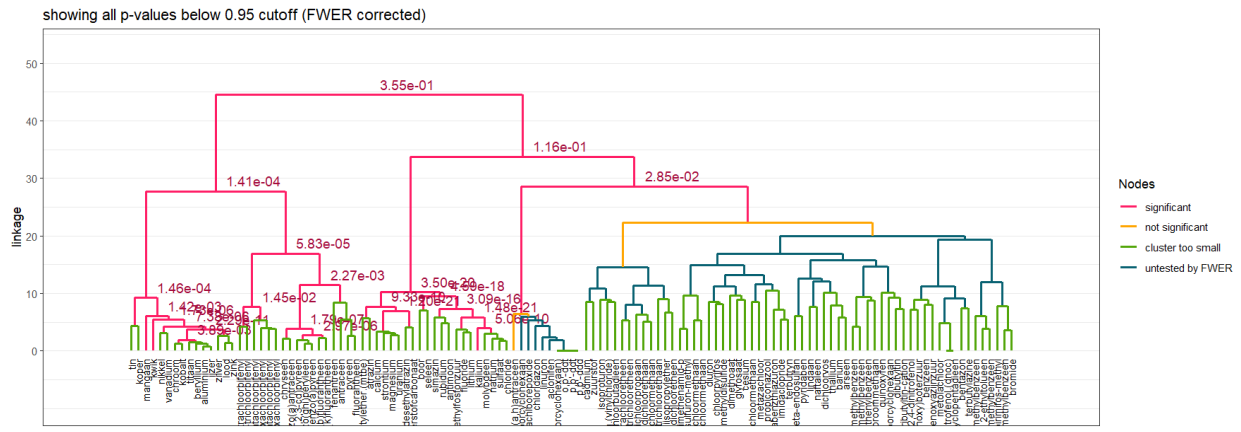


Figure Appendix III B. Significant clusters for location EYS with the R package Sigclust2.

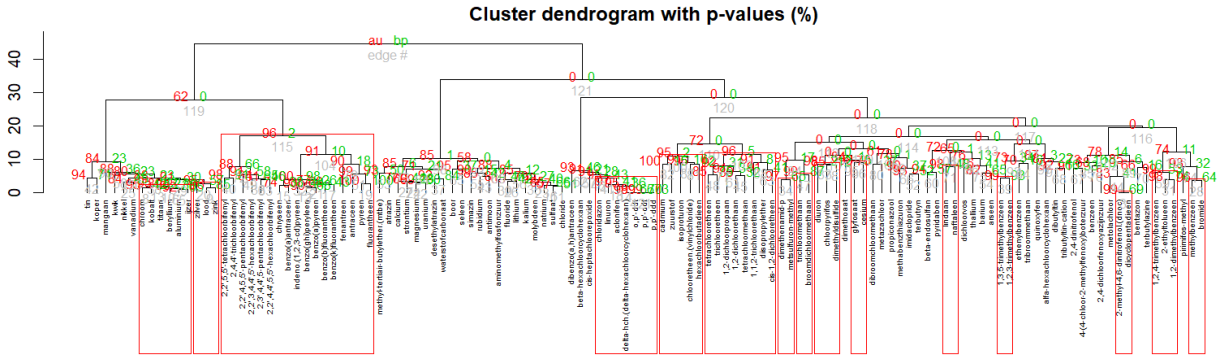


Figure Appendix III C. Significant clusters for the location EYS with the R package PVclust.

Table Appendix III A. Comparison of Cluster significance established visually, with the ClusSig method, PVclust, and Sigclust2.

Location Eysden	Visual assessment	ClusSig	Pvclust	Sigclust2
koper	cluster1			cluster1
tin	cluster1			cluster1
mangaan	cluster1	cluster1		cluster1
titaan	cluster1	cluster1	cluster2	cluster1
aluminium	cluster1	cluster1	cluster2	cluster1
ijzer	cluster1	cluster1	cluster2	cluster1
beryllium	cluster1	cluster1	cluster2	cluster1
chroom	cluster1	cluster1	cluster2	cluster1
kobalt	cluster1	cluster1	cluster2	cluster1
zink	cluster1	cluster1	cluster3	cluster1
lood	cluster1	cluster1	cluster3	cluster1
zilver	cluster1	cluster1	cluster3	cluster1

vanadium	cluster1	cluster1		cluster1
nikkel	cluster1	cluster1		cluster1
kwik	cluster1	cluster1		cluster1
2,2',3,4,4',5'-hexachloorbifenyl	cluster2	cluster2	cluster4	cluster2
2,3',4,4',5-pentachloorbifenyl	cluster2	cluster2	cluster4	cluster2
2,2',4,4',5,5'-hexachloorbifenyl	cluster2	cluster2	cluster4	cluster2
2,2',4,5,5'-pentachloorbifenyl	cluster2	cluster2	cluster4	cluster2
2,4,4'-trichloorbifenyl	cluster2	cluster2	cluster4	cluster2
2,2',5,5'-tetrachloorbifenyl	cluster2	cluster2	cluster4	cluster2
fenantreen	cluster3		cluster4	cluster3
antraceen	cluster3		cluster4	cluster3
pyreen	cluster3		cluster4	cluster3
fluorantheen	cluster3		cluster4	cluster3
benzo(a)pyreen	cluster3	cluster3	cluster4	cluster3
benzo(ghi)peryleen	cluster3	cluster3	cluster4	cluster3
indeno.(1,2,3-cd)pyreen	cluster3	cluster3	cluster4	cluster3
benzo(k)fluorantheen	cluster3	cluster3	cluster4	cluster3
benzo(b)fluorantheen	cluster3	cluster3	cluster4	cluster3
chryseen	cluster3	cluster3	cluster4	cluster3
benzo(a)antraceen	cluster3	cluster3	cluster4	cluster3
zuurstof			cluster5	
cadmium			cluster5	
chlooretheen.(vinylchloride)			cluster5	
hexachloorbutadieen			cluster5	
isoproturon			cluster5	
trichlooretheen	cluster4	cluster4	cluster6	
tetrachlooretheen	cluster4	cluster4	cluster6	
1,2-dichloorpropaan	cluster4	cluster4	cluster6	
tetrachloormethaan	cluster4	cluster4	cluster6	
1,2-dichloorethaan	cluster4	cluster4	cluster6	
1,1,2-trichloorethaan	cluster4		cluster6	
cis-1,2-dichlooretheen	cluster4		cluster6	
diisopropylether			cluster6	
terbutylazine				
bentazon				
2-methyl-4,6-dinitrofenol.(dnoc)		cluster5	cluster13	
dicyclopentadieen		cluster5	cluster13	
metolachloor		cluster5		
2,4-dichloorfenoxyazijnzuur				
benzeen				
4-(4-chloor-2-methylfenoxy)boterzuur	cluster5			
2,4-dinitrofenol	cluster5			
methylbenzeen	cluster5		cluster15	
bromide	cluster5		cluster15	
pirimifos-methyl	cluster5			
2-ethyltolueen	cluster5		cluster14	

1,2-dimethylbenzeen	cluster5		cluster14	
1,2,4-trimethylbenzeen	cluster5		cluster14	
lindaan			cluster11	
naftaleen			cluster11	
dichloorvos				
barium				
arseen				
thallium				
pyridaben				
beta-endosulfan				
1,2,3-trimethylbenzeen			cluster12	
1,3,5-trimethylbenzeen			cluster12	
ethenylbenzeen				
tribroommethaan				
quinoxyfen				
alfa-hexachloorcyclohexaan				
dibutyltin				
tributyltin-cation				
dimethenamid-p			cluster7	
metsulfuron-methyl			cluster7	
broomdichloormethaan			cluster8	
trichloormethaan			cluster8	
diuron			cluster9	
dimethyldisulfide			cluster9	
chloorpyrifos			cluster9	
dimethoaat				
glyfosaat			cluster10	
cesium			cluster10	
metazachloor				
dibroomchloormethaan				
propiconazool				
methabenzthiazuron				
terbutryn				
imidaclopride				
beta-hexachloorcyclohexaan	cluster6	cluster6		cluster6
cis-heptachloorepoxide	cluster6	cluster6		cluster6
p,p'-ddt	cluster6	cluster6		cluster6
p,p'-ddd	cluster6	cluster6		cluster6
o,p'-ddt	cluster6	cluster6		cluster6
delta-hch.(delta-hexachloorcyclohexaan)	cluster6	cluster6		cluster6
aclonifen	cluster6	cluster6		cluster6
linuron	cluster6	cluster6		cluster6
chloridazon	cluster6	cluster6		cluster6
dibenzo(a,h)antraceen	cluster6	cluster6		cluster6
atrazin	cluster7	cluster7		cluster4
methyl-tertiair-butylether.(mtbe)	cluster7	cluster7		cluster4

desethylatrazin	cluster7	cluster7		cluster4
waterstofcarbonaat	cluster7	cluster7		cluster4
calcium	cluster7	cluster7		cluster4
magnesium	cluster7	cluster7		cluster4
uranium	cluster7	cluster7		cluster4
strontium	cluster7	cluster7		cluster4
boor	cluster7			cluster5
seleen	cluster7			cluster5
rubidium	cluster7	cluster8		cluster5
antimoon	cluster7	cluster8		cluster5
simazin	cluster7	cluster8		cluster5
chloride	cluster7	cluster8		cluster5
sulfaat	cluster7	cluster8		cluster5
natrium	cluster7	cluster8		cluster5
molybdeen	cluster7	cluster8		cluster5
kalium	cluster7	cluster8		cluster5
aminomethylfosfonzuur	cluster7	cluster8		cluster5
fluoride	cluster7	cluster8		cluster5
lithium	cluster7	cluster8		cluster5

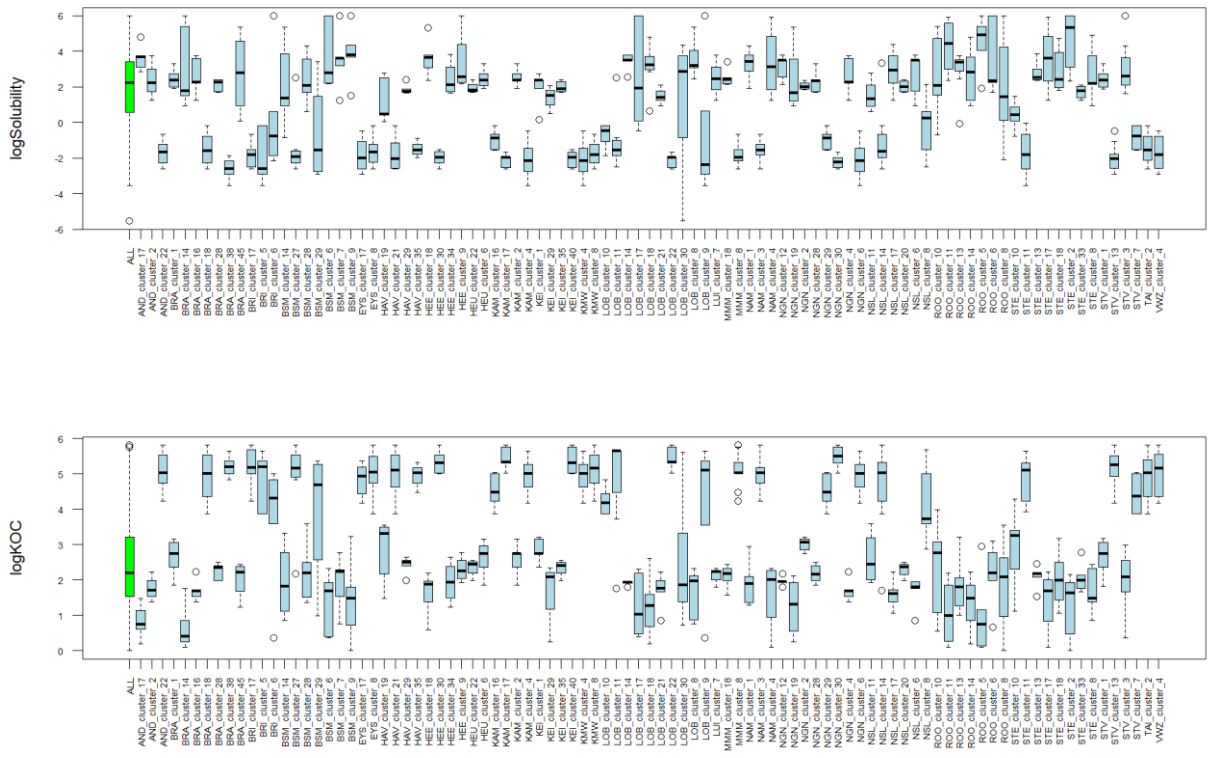
IV Substance properties and environmental conditions per cluster

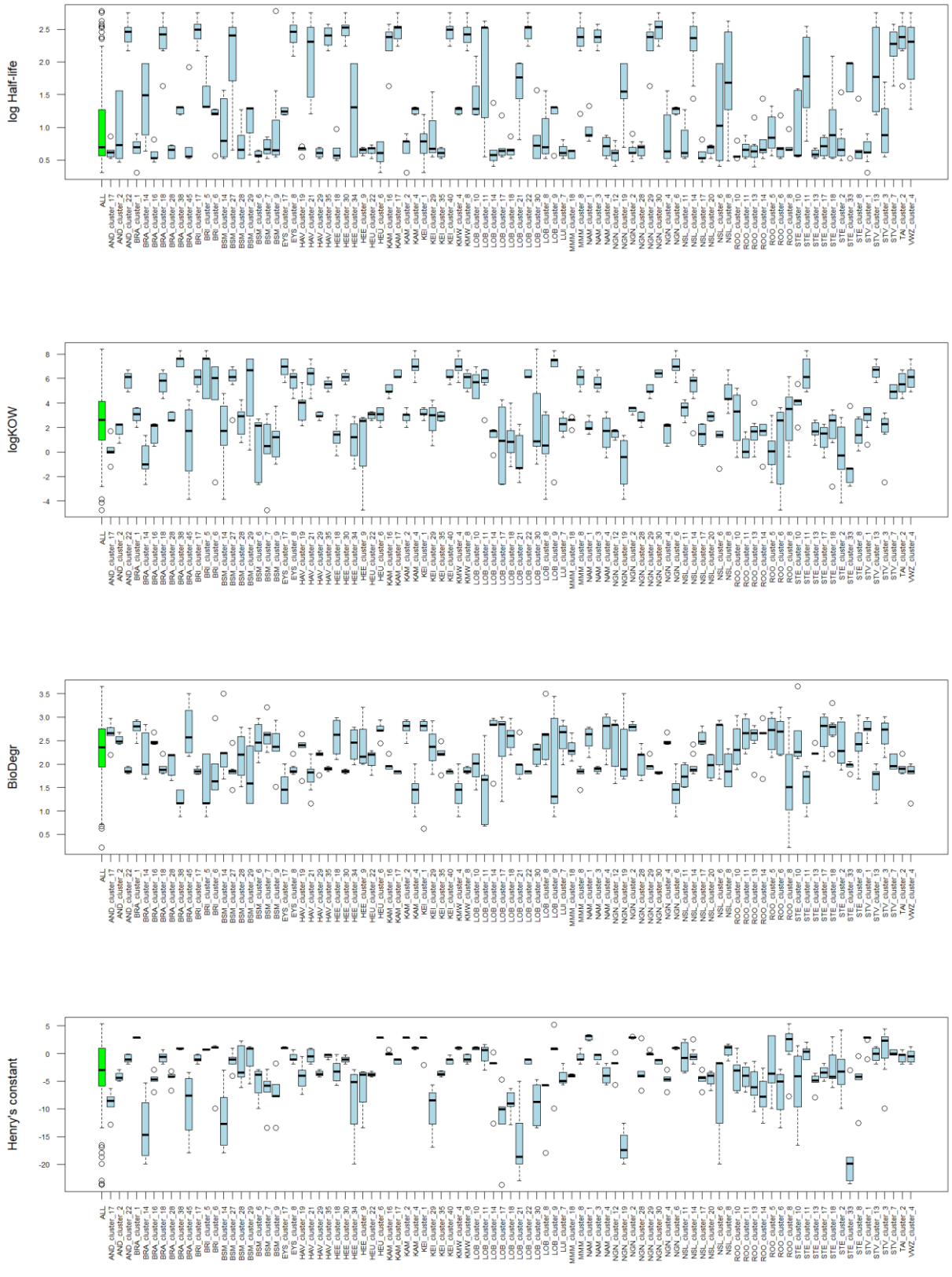
In this appendix, the property values for the substances in clusters per location are visualised (below Table Appendix IV). All clusters with less than 5 substances were removed. Also inorganic substances were omitted because the models used could not predict the substance properties for inorganic substances. Boxplots of environmental conditions with high concentrations of substances in clusters were made separately, one for the environmental values derived for the Rhine, one for the values from the Meuse.

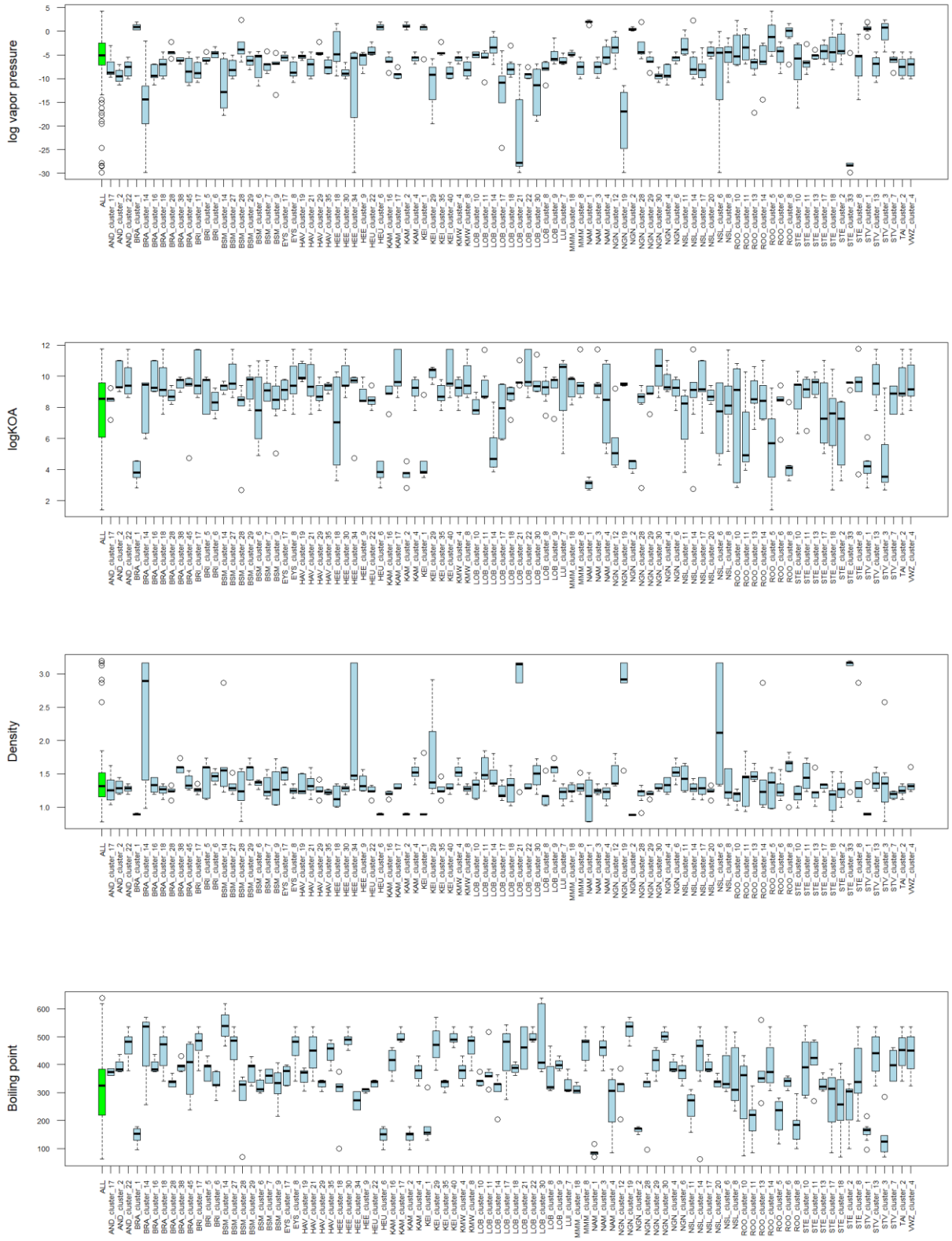
Table Appendix IV A. Examples of relation of properties and conditions to the transport and fate of substances. Adapted from: Agency for Toxic Substances and Disease Registry

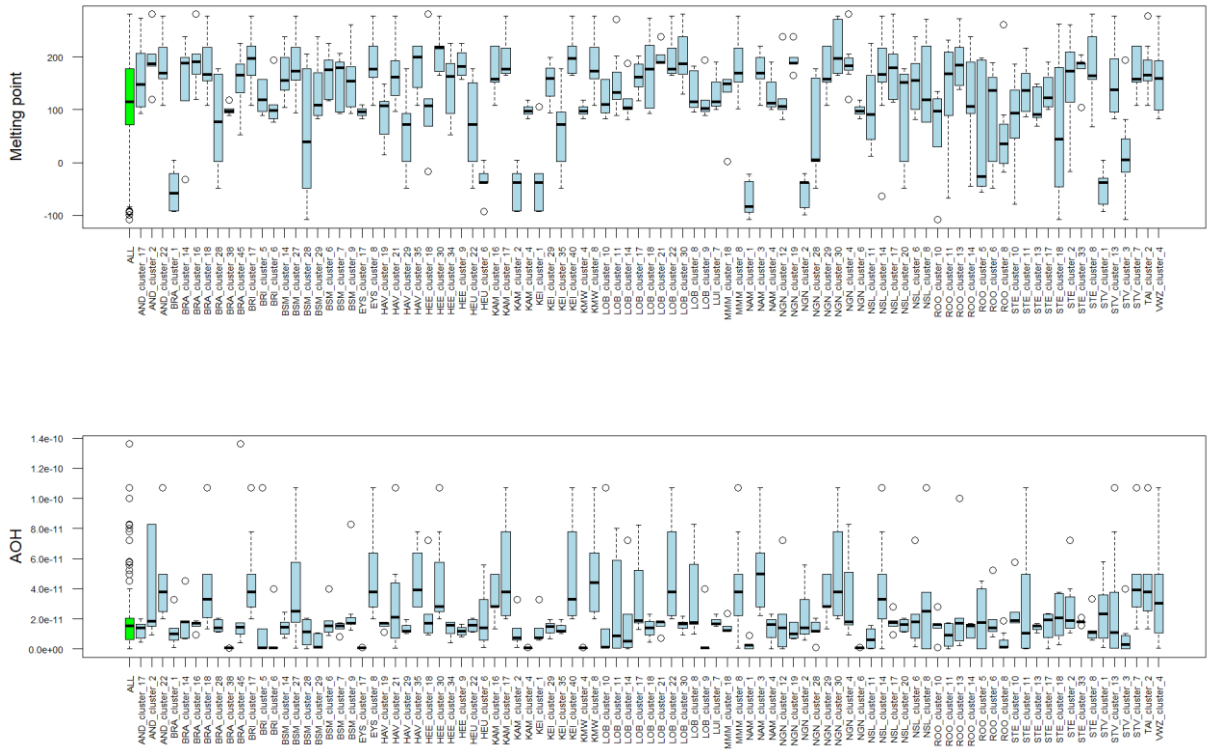
Substance property	Definition	Explanation
Water solubility	The maximum concentration of a chemical that dissolves in a given amount of pure water.	Physicochemical parameters, such as salinity and pH, can influence a chemical's solubility in water, which, in turn, also affects its dissolved concentration in water. Solubility provides an important indication of a contaminant's mobility in the aquatic environment, and its ability to reach drinking water sources such as groundwater.
Density of liquid	A liquid's mass per volume.	For liquids (typically organic solvents) that are immiscible in water, density plays a critical role. In groundwater, liquids with a higher density than water may penetrate and preferentially settle to the base of an aquifer, while less dense liquids will float.
Vapor pressure	A measure of the volatility of a chemical in its pure state.	Vapor pressure largely determines how quickly contaminants will evaporate from surface soils or water bodies into the air. Contaminants with higher vapor pressures will evaporate more readily.
Henry's Law Constant	A measure of the tendency for a chemical to pass from an aqueous solution to the vapor phase.	A high Henry's Law Constant corresponds to a greater tendency for a chemical to volatilize to air. It is a function of molecular weight, solubility, and vapor pressure.
Organic carbon partition coefficient (Koc) ('Adsorption	The sorption affinity of a chemical for organic carbon and consequently the tendency for	A high Koc indicates a stronger binding affinity to organic matter. In soil and sediment this may results in reduced

coefficient')	compounds to be adsorbed to soil and sediment.	mobility, and thus, less of the chemical is available to move into groundwater or surface water.
Octanol-water partition coefficient (Kow)	The ratio between the concentration of a substance in octanol and water in a biphasic octanol/ water system	Provides a measure of the polarity of a substance and its ability to partition to water. A low Kow is indicative of a polar substance. A polar substance is expected to be more mobile in the aquatic environment and typically more challenging to remove from water. In contrast, hydrophobic (high Kow) substances tend to accumulate on solid particles and as a result occur at lower concentrations in the aqueous phase.
Half-life	The time it takes to reduce an environmental concentration of a chemical by half due to chemical physical or biological processes.	Media-specific half-life provides a relative measure of how persistent a contaminant might be in a particular environmental medium by processes that generally involve reactions like hydrolysis, oxidation / reduction, and photolysis

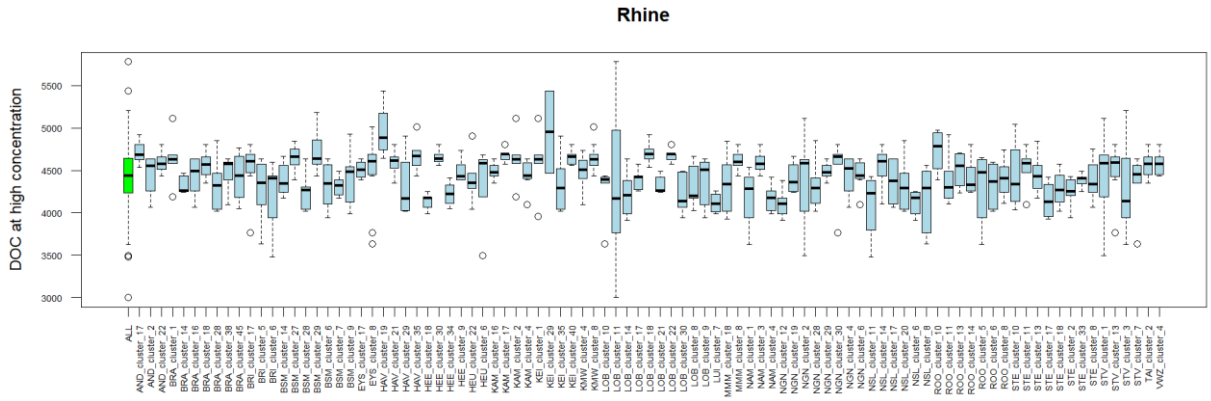


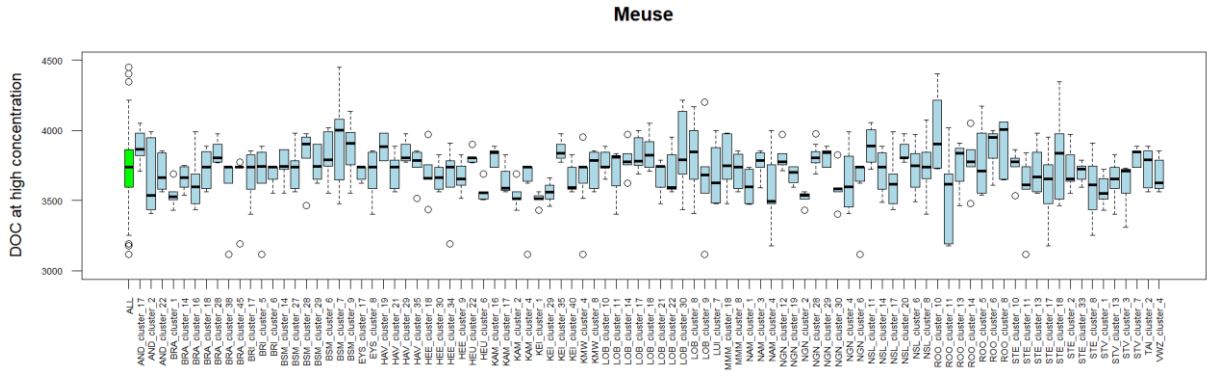

log Half-life
logKOW
BioDegr
Henry's constant

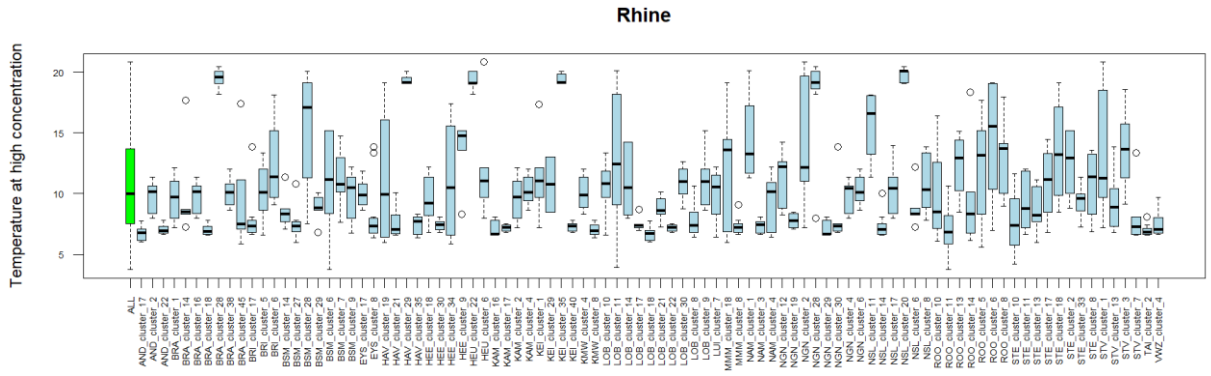

log vapor pressure
logKOA
Density
Boiling point

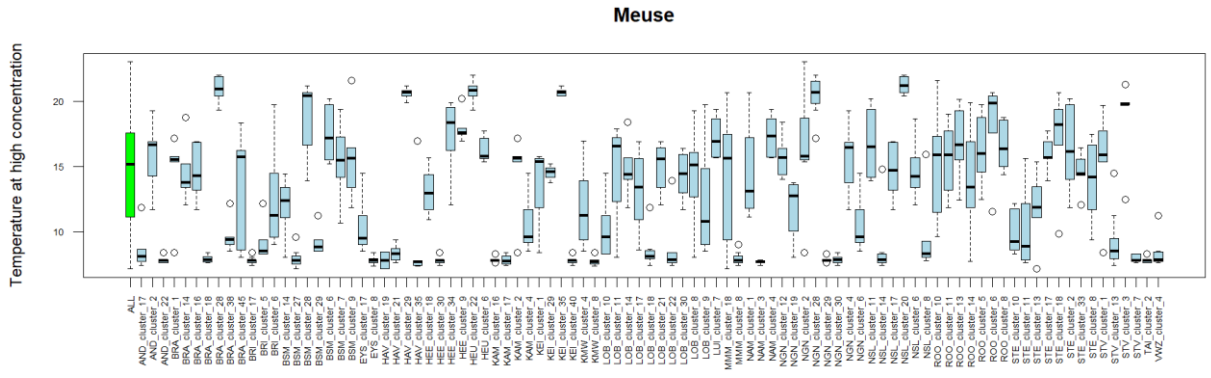


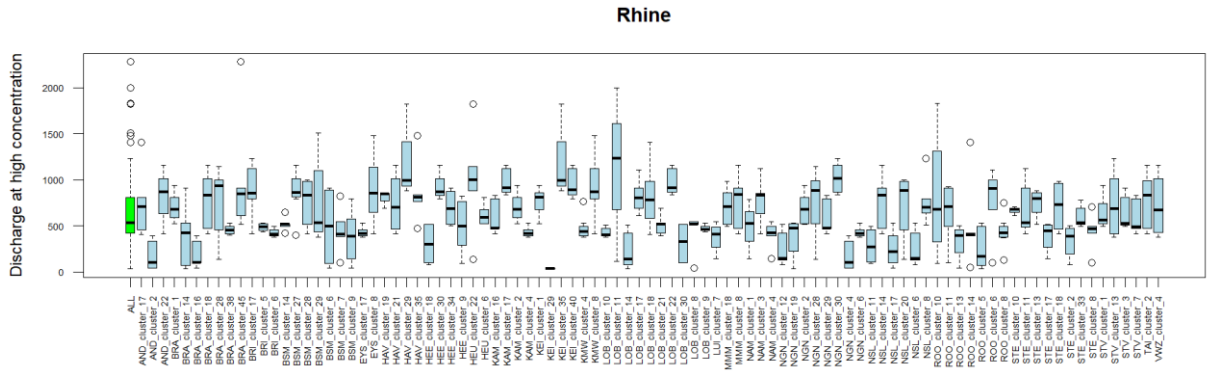
Note: Below, clusters of both locations (Rhine and Meuse) are shown for each environmental condition. This gives extra information and an opportunity to check whether the substances in a cluster from one river system have similar response to environmental conditions in the other river system. 'High concentration' (see y-axis) refers to the top 10 percent of concentrations of substances in the cluster for either Rhine or Meuse.

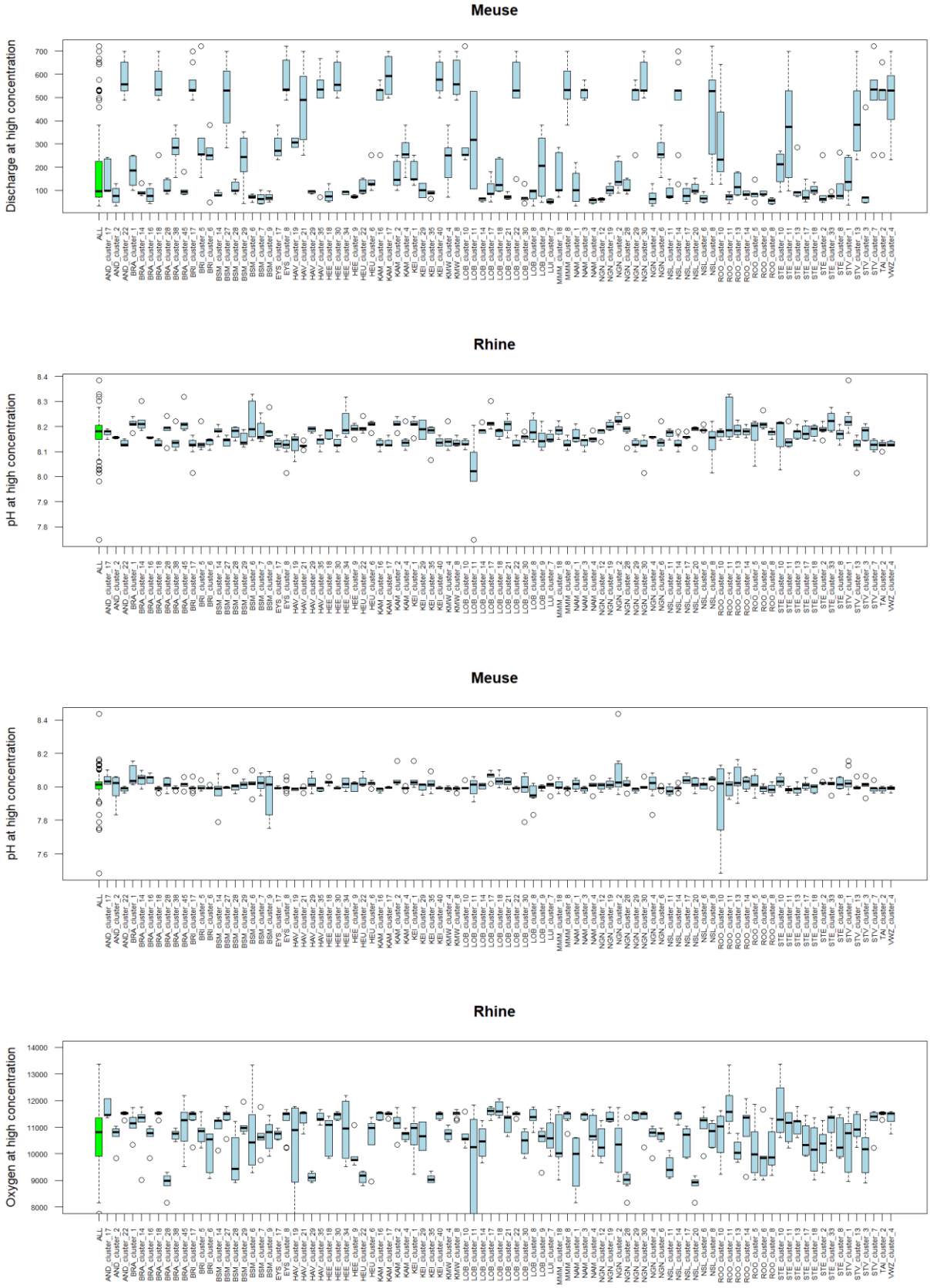


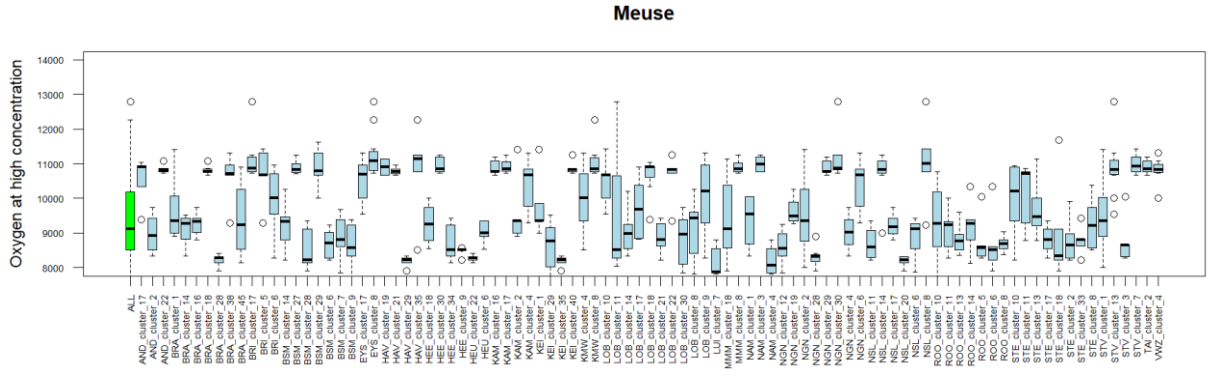

Meuse
DOC at high concentration
4500
4000
3500
3000

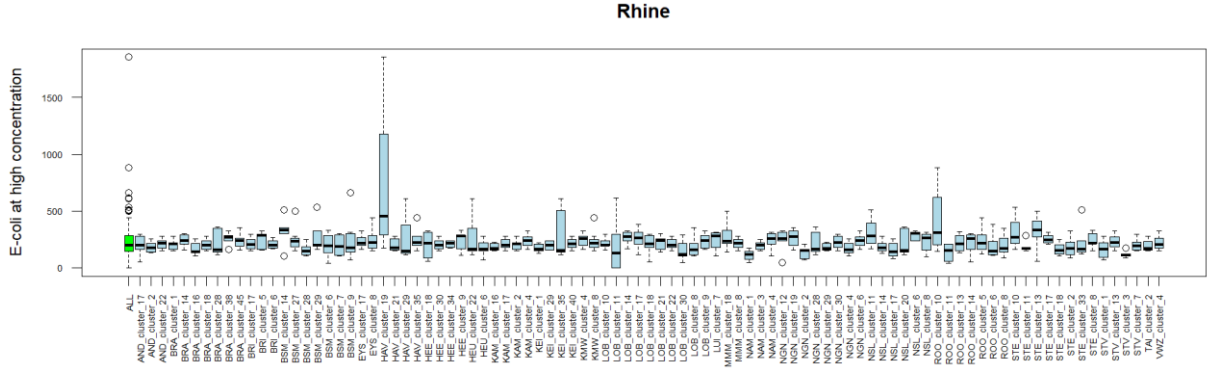

Rhine
Temperature at high concentration
20
15
10
5

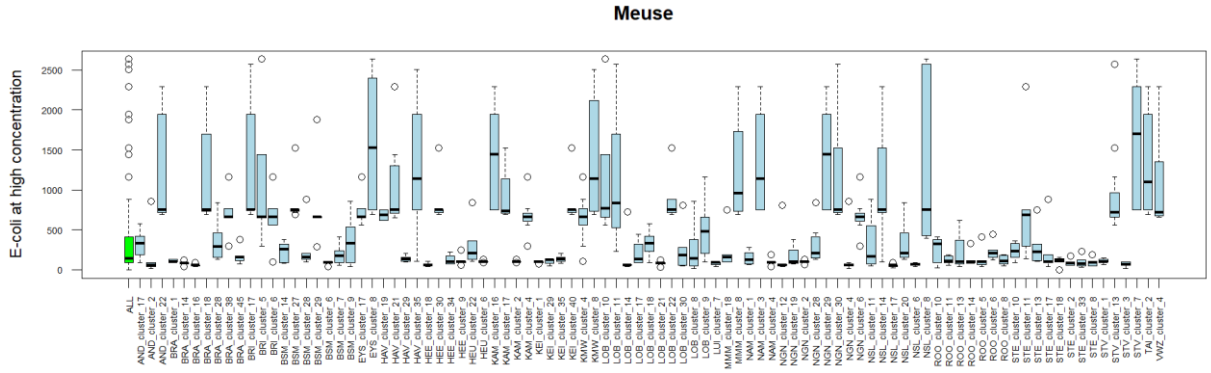

Meuse
Temperature at high concentration
20
15
10

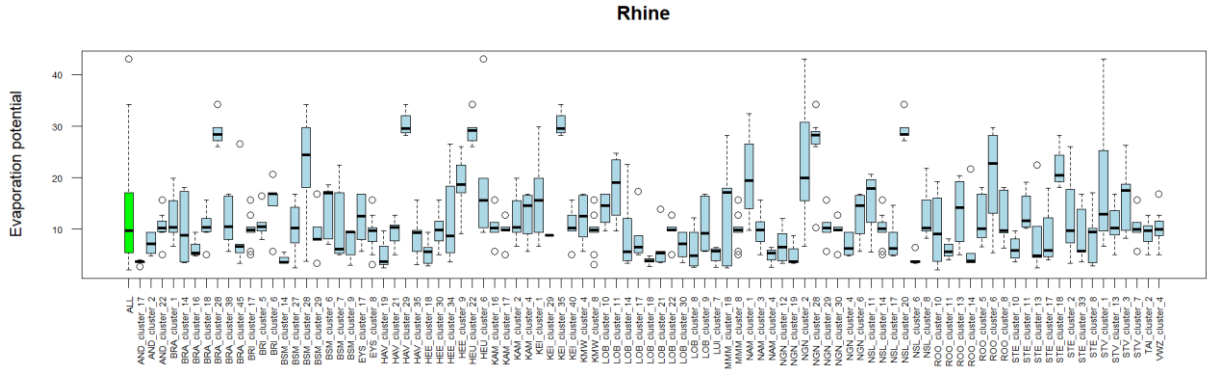

Rhine
Discharge at high concentration
2000
1500
1000
500
0

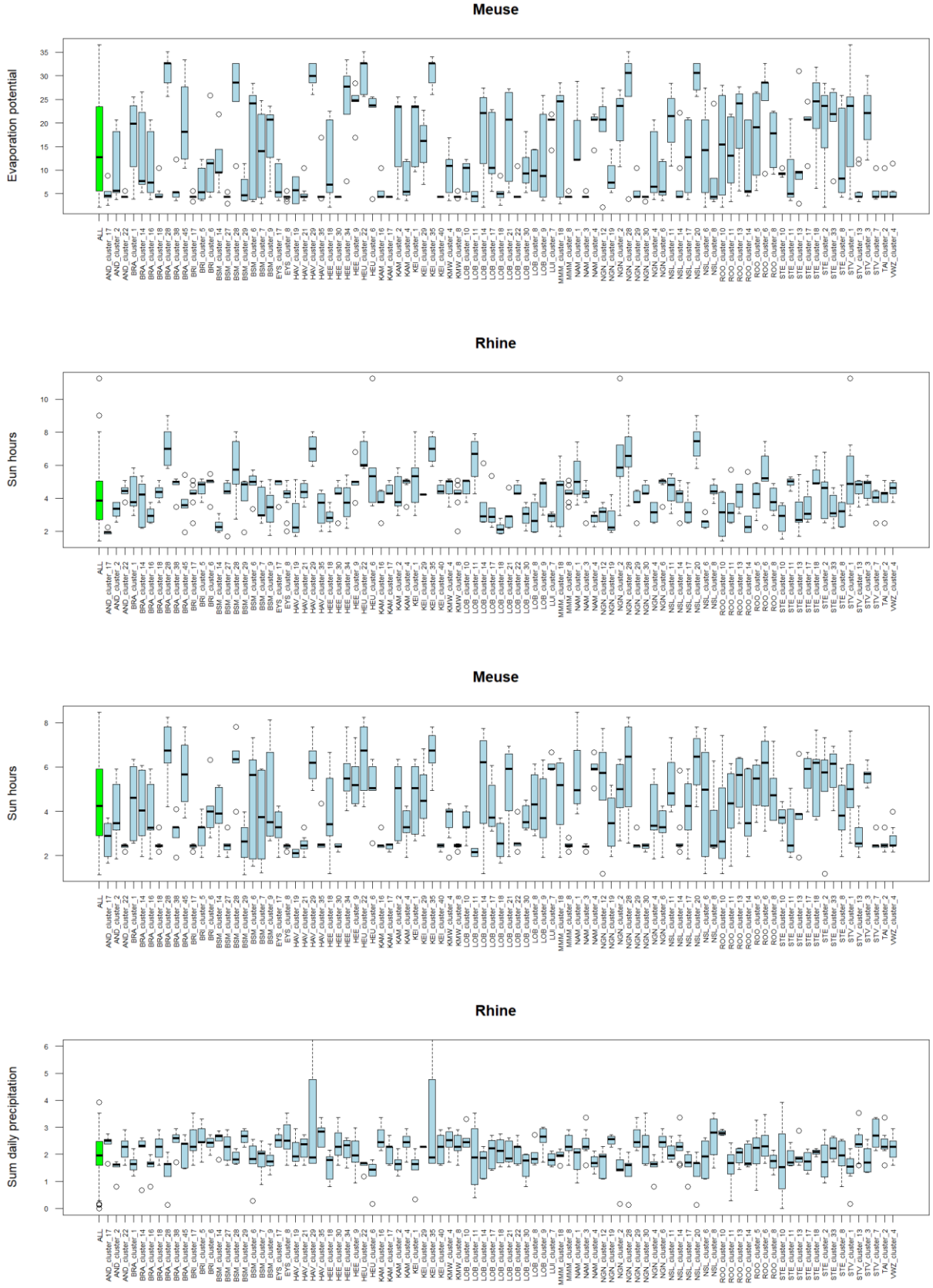

Meuse
Discharge at high concentration
Rhine
pH at high concentration
Meuse
pH at high concentration
Rhine
Oxygen at high concentration

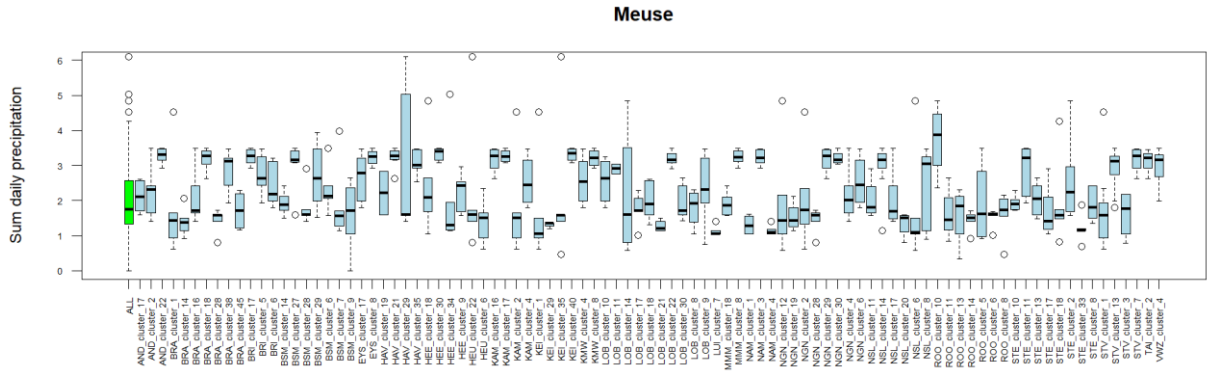

Meuse
Oxygen at high concentration
14000
13000
12000
11000
10000
9000
8000


Rhine
E-coli at high concentration
1500
1000
500
0


Meuse
E-coli at high concentration
2500
2000
1500
1000
500
0


Rhine
Evaporation potential
40
30
20
10


Meuse
Evaporation potential
Rhine
Sun hours
Meuse
Sun hours
Rhine
Sum daily precipitation


Meuse
Sum daily precipitation
0
1
2
3
4
5
6
ALL
AND_cluster_17
AND_cluster_21
AND_cluster_22
BRA_cluster_1
BRA_cluster_14
BRA_cluster_16
BRA_cluster_18
BRA_cluster_28
BRA_cluster_38
BRA_cluster_45
BRI_cluster_17
BRI_cluster_5
BRI_cluster_6
BSM_cluster_14
BSM_cluster_27
BSM_cluster_28
BSM_cluster_29
BSM_cluster_6
BSM_cluster_7
BSM_cluster_9
EYS_cluster_17
EYS_cluster_8
HAV_cluster_19
HAV_cluster_21
HAV_cluster_29
HAV_cluster_35
HEE_cluster_18
HEE_cluster_30
HEE_cluster_34
HEE_cluster_9
HEU_cluster_22
HEU_cluster_6
KAN_cluster_16
KAN_cluster_17
KAN_cluster_2
KAN_cluster_4
KEI_cluster_1
KEI_cluster_29
KEI_cluster_35
KEI_cluster_40
KMW_cluster_4
KMW_cluster_8
LOB_cluster_10
LOB_cluster_11
LOB_cluster_14
LOB_cluster_17
LOB_cluster_18
LOB_cluster_21
LOB_cluster_22
LOB_cluster_30
LOB_cluster_8
LOB_cluster_9
LUI_cluster_7
MAA_cluster_18
MAA_cluster_8
NAM_cluster_1
NAM_cluster_3
NAM_cluster_4
NGN_cluster_12
NGN_cluster_19
NGN_cluster_2
NGN_cluster_28
NGN_cluster_29
NGN_cluster_30
NGN_cluster_4
NGN_cluster_6
NSL_cluster_11
NSL_cluster_14
NSL_cluster_17
NSL_cluster_20
NSL_cluster_6
NSL_cluster_8
ROO_cluster_10
ROO_cluster_11
ROO_cluster_13
ROO_cluster_14
ROO_cluster_5
ROO_cluster_6
ROO_cluster_8
STE_cluster_10
STE_cluster_11
STE_cluster_13
STE_cluster_17
STE_cluster_18
STE_cluster_2
STE_cluster_33
STE_cluster_8
STV_cluster_1
STV_cluster_13
STV_cluster_3
STV_cluster_7
TAL_cluster_2
VAK_cluster_4

V Groundwater monitoring data

For a proof of concept of *spatial analysis* with overlapping reference lists we selected a groundwater quality set provided by the Dutch provinces containing micropollutants concentrations for a single period across the Netherlands. This dataset is available on the 'Waterkwaliteitsportaal', however we used an in-house version with processed data resulting from a previous project.

The groundwater data involves data collected between 2016 and 2018. This is summarized into a single measurement. Because groundwater is less variable in time, the most recent sample was selected per location. For most locations the same parameters were measured, so there are only a few missing values. On the contrary, because groundwater is in general less polluted than surface water, a relatively high number of measurements were below Reporting Limit.

Table Appendix V A. Data characterization of the groundwater dataset.

Groundwater, data of the Provinces
505 parameters
Single sampling (between 2016-2018)
564 locations

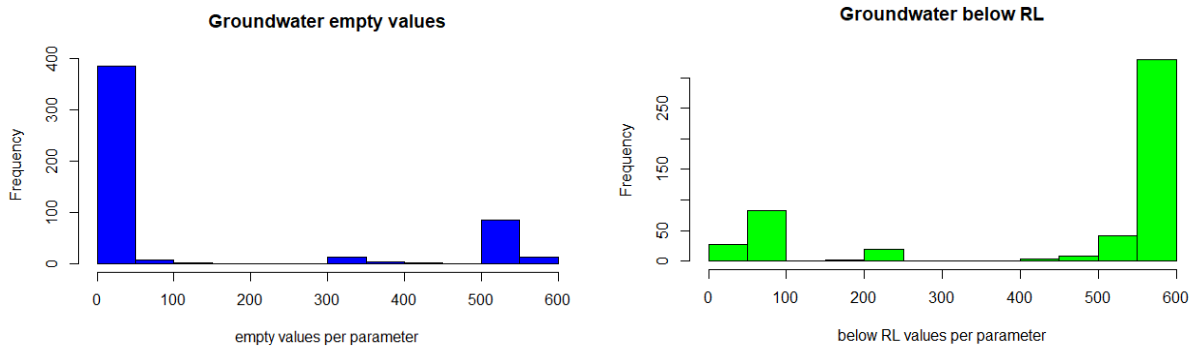


Figure Appendix V A. The 'frequency' on the y-axis denotes the number of parameters. Left: missing values per parameter. Right: values below RL per parameter. The maximum missing values / below RL per parameter is 564 (the amount of locations).

When parameters without *any* measurements above RL were removed (this includes missing values), in total 176 parameters from 505 were left for the analyses (Table a).

Associating measurement data to sources of pollution

Groundwater typically contains less microcontaminants than surface water. This means there are less chemicals detected in groundwater. Indeed, in the groundwater dataset, many measurements have only a few chemicals above the reporting limit. The combined occurrence of substances is quite unique for every location. There are, however, some substances in the groundwater dataset that were detected in many locations, and locations in which many substances (up to 27) are detected. Still, some potential reference substance lists can be linked to samples, and

plotted on a map. The linkage is done by performing the hypergeometric test and selecting samples that significantly overlap (with at least two parameters) with a reference substance list of choice. Figure b shows spatially explicit maps of reference substance lists that significantly overlap with measurement data. Many other reference substance lists are significant in a number of measurement data (from a lot of locations to just incidentally). We choose only four as an example.

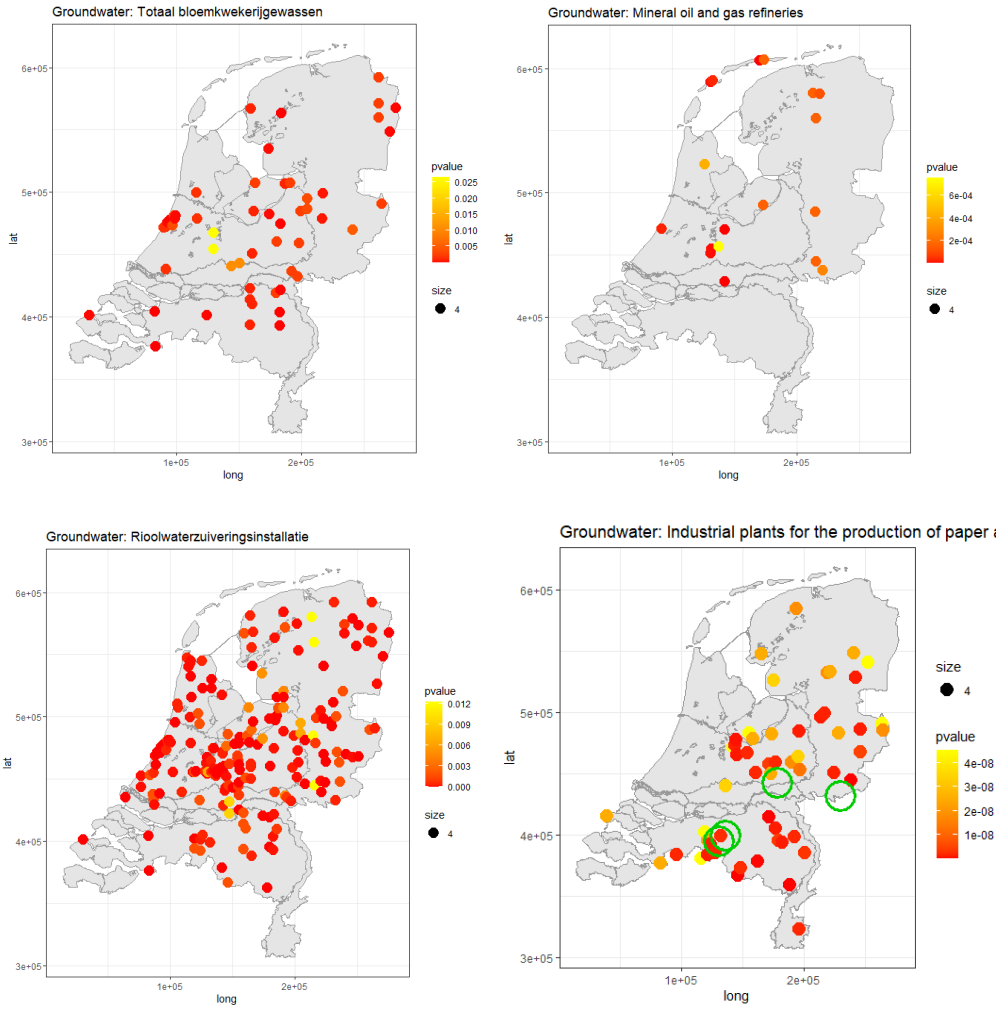


Figure Appendix V B. Examples of maps with significant overlap between detected substances in samples and reference substance lists. In the figure on the bottom-right, four known paper production plants are indicated in green circles. An exhaustive search for paper production plants was not done.

When comparing the maps in Figure b, a difference in spatial distribution can be observed in samples with overlap for different reference substance lists. This can lead to hypotheses on the origin of some of the substances in the measurement sample. It could be that all of the samples that overlap with the reference substance list 'Totaal bloemkwekerijgewassen' ('Total flower nursery crops') are influenced by this particular type of agriculture. Similarly this would apply to examples in Figure b that depict samples that are overlapping in the other reference substance lists, e.g. samples influenced by mineral oil and gas refinery – related activity, influenced by sewage treatment plant effluent, influenced by paper industry- related activity. Before this type of environmental forensics analyses can be done, several adjustments are still needed. It is important to incorporate how (ground)water flows and what distance these substances may travel in time, in other words what is the age of the GW, so what period of emission diooes it represent. In addition this also requires data on historical activities to enable the explanation of results, for such

analysis 'gebiedsdossiers' can be of value. Similarly, substances that are rarely found but have a reason (high degradability, high volatility, high adsorption) need to be recognized. For naturally occurring substances (e.g., metals) finding a significant overlap is now still less informative because they occur nearly everywhere. With such substances either a more stringent selection criterium (i.e., p-value) is necessary, or these substances can be considered only if they exceed their background value. This will have to be addressed in a follow-up of this research. A validation of the detected sources of influence is also very valuable to do. This can be achieved by linking samples to actual known sources or activities. In this way, the validity of the overlap can be checked for some of the reference substance lists that have emissions via physical locations.

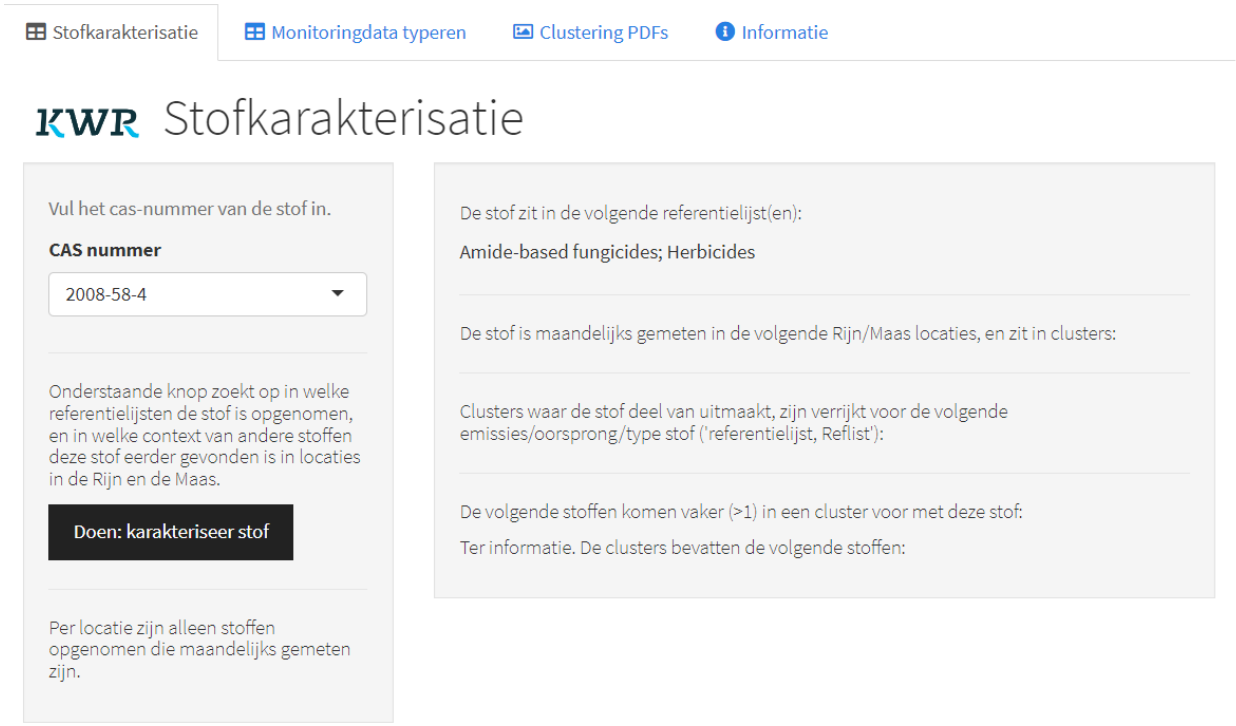
Of course, for a single sample, several significant overlaps between the measured data and reference substance lists can be detected. This provides an overview of the potential land uses and activities that may impact a given sample. Table b lists some examples of significant overlap of detected substances for groundwater samples with substances in the reference substance lists. Aside from these examples, there are many locations *without* a single significant overlap. These samples are typically characterized by a very limited number of substances measured above the reporting limit.

Table Appendix V B. Examples of locations with significant overlap with different reference substance lists. For these analyses, nutrients and generic parameters were not included.

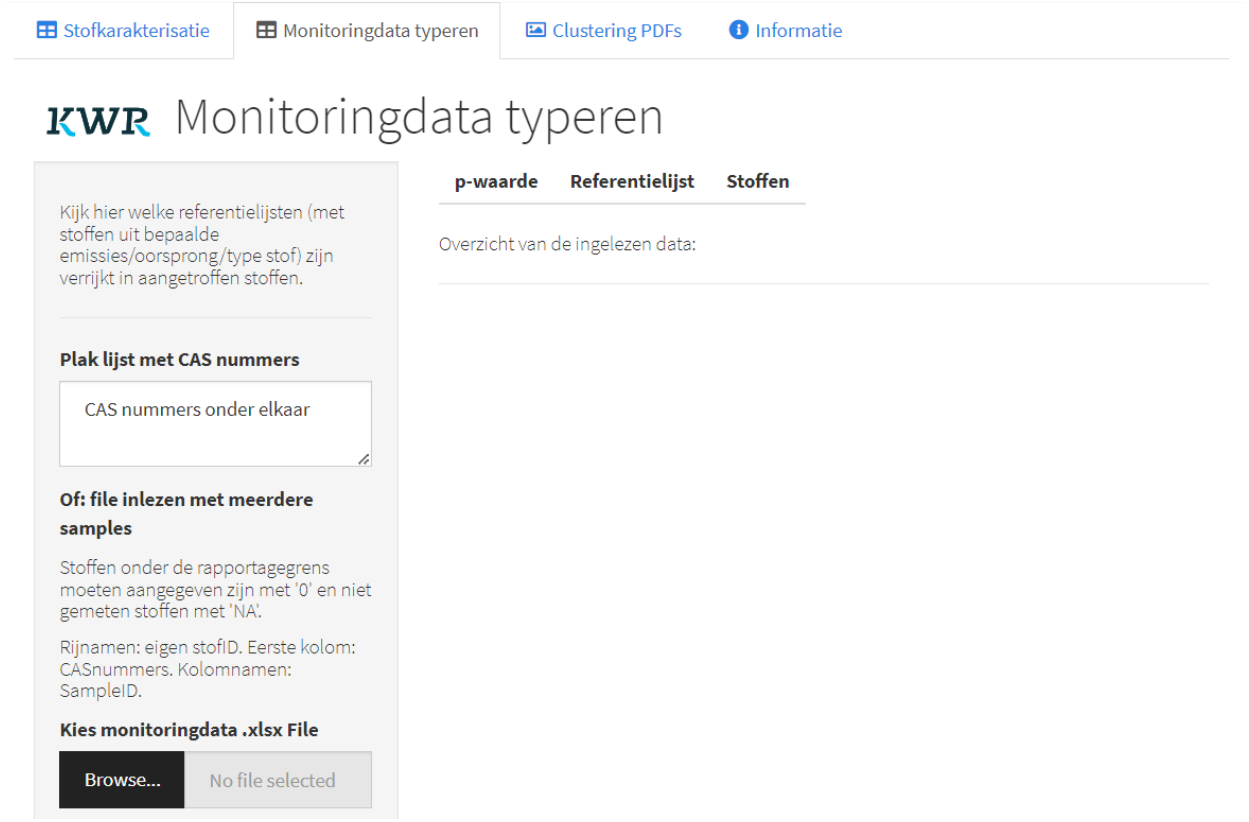
Province	Ground water body	CAS Substances overlapping	Associated reference lists, significant overlap
NOORD-HOLLAND	NLGW0016	15950-66-0#933-78-8#933-75-5	Industrial chemicals (with phenols)
		335-67-1#1763-23-1	Industrial chemicals (with per- and polyfluoroalkyl substances (PFAS))
NOORD-BRABANT	NLGW0006	2008-58-4#3984-14-3	Amide-based fungicides
		2008-58-4#6339-19-1#17254-80-7	Herbicides
UTRECHT	NLGW0012	25057-89-0#1698-60-8	Seed onions/Total floricultural crops/Foot and plant onions/Daffodils
		1698-60-8#134-62-3#93-65-2#80-05-7#15307-86-5#25812-30-0	Waste water treatment plant
		57-68-1#127-79-7#144-83-2	Antibiotics based on sulphonamides
		78-40-0#126-71-6	Flame retardants/industrial solvents
		2008-58-4#3984-14-3	Amide-based fungicides
		2008-58-4#25057-89-0#1698-60-8	Herbicides
		15307-86-5#60-80-0#125-33-7	Pain killers and fever reducing substances
		60166-93-0#125-33-7	Domestic wastewater
		15307-86-5#60-80-0	Aerobic/anaerobic conditions

VI Screenshots of the 'shiny R' app

An R shiny app was developed to work with the many clusters found in this project in the Rhine and Meuse. In this Appendix, screenshots of the sheets in the app are shown, with a short explanation of the functionality.

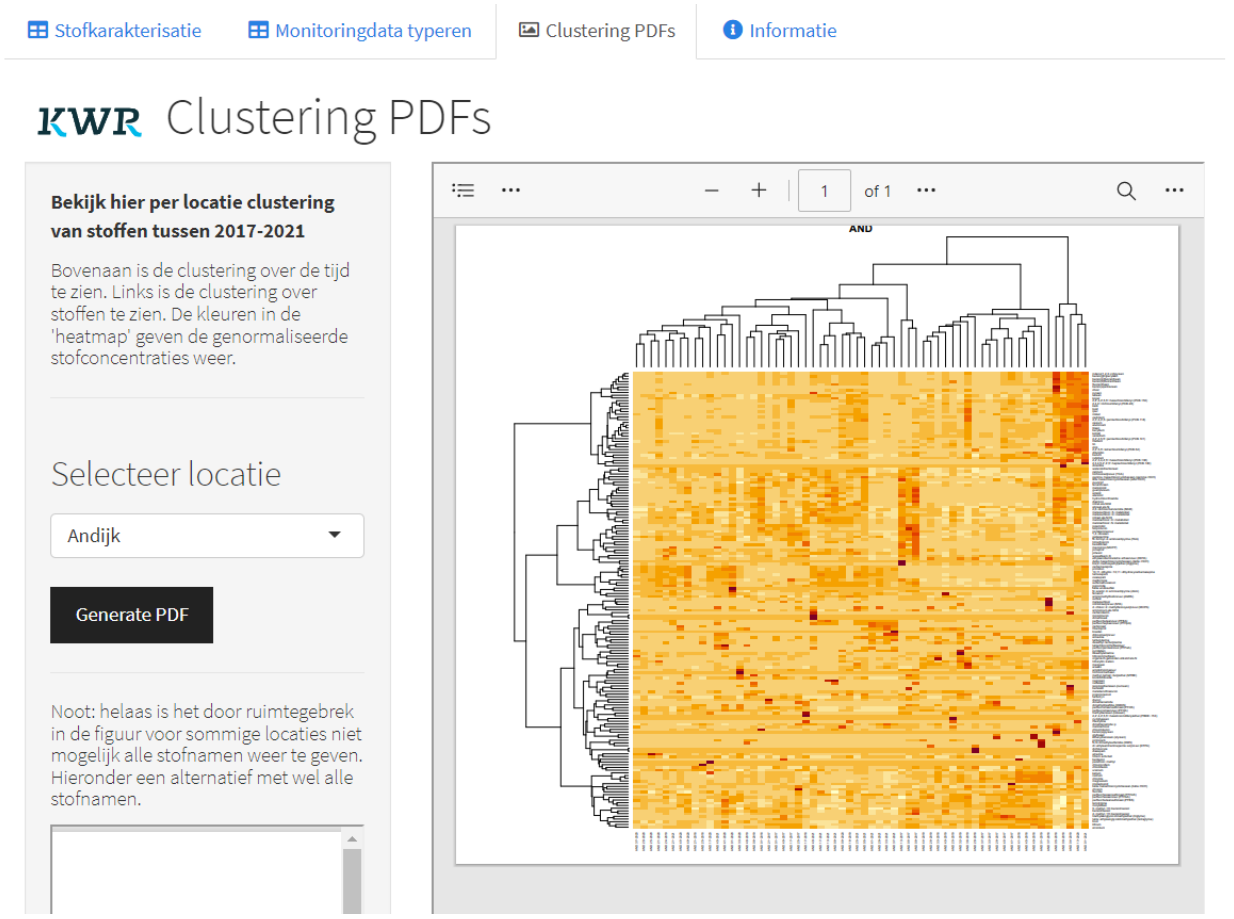


Sheet 1. For a CAS number of interest (left panel) some statistics are shown (right panel). The statistics are, in order from top to bottom: the Reference lists that the substance is a member of, the clusterID's that contain the substance of interest, the overlap with reference lists for the clusters with the substance of interest, which substances occur together more often in clusters with the substance of interest, an overview of all substances in clusters with the substance of interest.

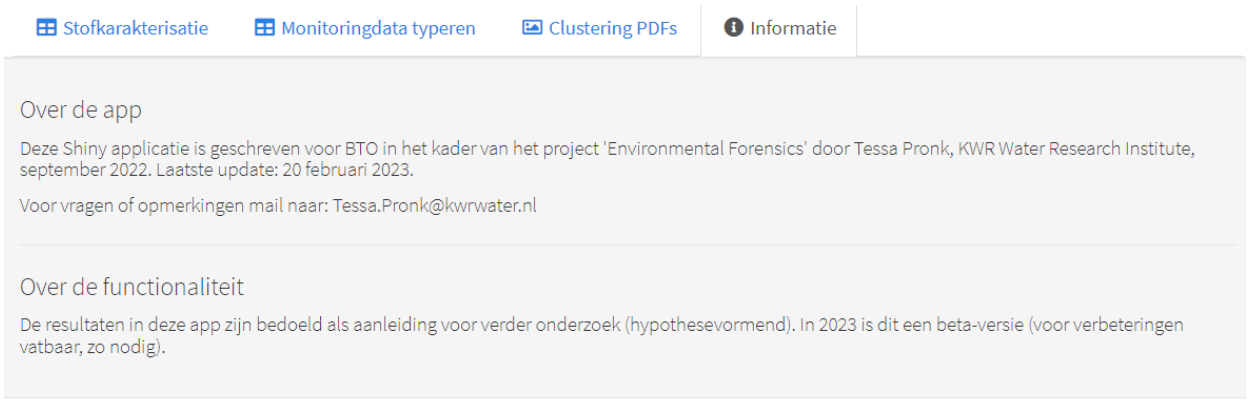


Sheet 2, top part. In the left panel, a list of CAS numbers can be pasted. For instance, detected in a particular sample or substances in a cluster. Then, the significance of the overlap with Reference lists is calculated. This is shown in the right panel. As an alternative, a local file with samples can be loaded. This is explained further below.

Sheet 2, lower part. If a file is uploaded, the sample of interest can be selected in the left panel. For that sample, the significance of the overlap with Reference lists is calculated. This is shown in the right panel. For a sample from a file, missing substances can be suggested. These are substances that are not in the file, but occur together in clusters in locations in Meuse and Rhine. A file that can be used for upload can be found in the data package (https://doi.org/10.5281/zenodo.8220952, 2023).



Sheet 3. Show the clustering results per location. In the right panel the dendrogram with heatmap is shown. There is an option to zoom in. In the left lower panel a dendrogram is shown that has all substance names.



Sheet 4. The information sheet with information on the version and intended use of the R shiny app.

VII Literature search of Environmental Forensics applications

Besides techniques and resources as discussed in this report, there are others that can be used, for particular purposes. Below is a literature search for used clustering techniques and for reference substances.

Literature on clustering techniques for monitoring data

The following techniques have (among others) been applied for clustering substances. Cluster analysis (CA) is an unsupervised pattern recognition method commonly used to group variables and observations. CA has been used to group sampling sites showing similar PAHs fingerprints into clusters to explain the variations between sites (Dahle et al., 2003; Savinov et al., 2000) and to identify sources of PAHs by grouping PAHs having similar characteristics (Kavouras et al., 2001; Liu et al., 2009). This approach was also used in combination with principal Component Analysis (PCA) and forensic tools (e.g., substance ratios, speciation) for the identification of pollution sources in estuarine areas linked to zinc smelting, coaly particles and waste disposal (Baragaño et al., 2022).

Discriminant analysis (DA) offers statistical classification of samples with prior knowledge of membership of objects to particular clusters (such as spatial or temporal grouping of a sample is known from its sampling sites or time). It is used to confirm the groups found by means of CA. In addition, DA helps in grouping the samples sharing the common properties (Al-Odaini et al., 2012; Kannel et al., 2007; Osman et al., 2012; Singh et al., 2005).

Principal Component Analysis (PCA) is used to reduce the number of variables and to detect structural relationships among the variables. For instance, PCA has been used to detect relationships among variables for possible source identification of PAHs in air, sediment, biota and soil (Harrison et al., 1996, Larsen and Baker, 2003, Luo et al., 2006, Luo et al., 2008, Pies et al., 2008, Gaspare et al., 2009). PCA has also been applied together with molecular indices for identification of sources of PAHs in complex environmental samples (Luo et al., 2006, Luo et al., 2008, Zuo et al., 2007, Pies et al., 2008, Liu et al., 2009).

Factor scores from PCA coupled with multiple linear regression (APCS/MLR) is a popular technique for source apportionment of PAHs in environmental matrices (Harrison et al., 1996; Kavouras et al., 2001; Larsen and Baker, 2003; Wang et al., 2010). The advantage of APCS/MLR is that it does not require prior knowledge on input of source emission to calculate source contributions (Larsen and Baker, 2003; Liu et al., 2009).

Polytopic vector analysis (PVA), a multivariate technique based on a linear mixing model, was used to identify a dioxin dechlorination fingerprint indicative of biotic/abiotic transformations in field samples of sediments (Barabás et al., 2004). PVA was also applied in combination with t-Distributed Stochastic Neighbor Embedding (t-SNE) to identify potential point source signatures in PAHs contaminated sediments (Jordan et al., 2021).

Random Forest (RF) is a more recently used machine learning method which allows to understand the individual role and the combined effect of explanatory variables. This approach has been applied to monitoring data to develop predictive models useful for investigating pollution sources in groundwater (Bindal and Singh, 2019; Rodriguez-Galiano et al., 2014), and biomonitoring parameters for the analysis of the biological impacts of multiple pressures in aquatic ecosystems (Feld et al., 2016). Variations of RF (e.g., extremely-randomized trees) (Geurts et al., 2006) can also be used to create decision tree methods which classify data by creating a network of choices based on the magnitudes of features, which have been applied for source allocation of PFAS (Kibbey et al., 2020).

Several studies have used logistic regression models (LRM) (Lee et al., 2009; Zhang et al., 2012) to assess the likelihood of As contamination greater than the predefined limit of 10 µg/L by using limited As data points along with auxiliary independent variables, such as geology, topography, and soil properties. A few studies used linear regression (LR) (Zhang et al., 2013), principal component regression (PCR) (Luo et al., 2012), Bayesian modeling (Cha et al., 2016) and artificial neural network (ANN)-based regression (Bonelli et al., 2017; Cho et al., 2011) for As prediction in groundwater and soil.

Literature on the use of indicator substances

Most organic micropollutants do not naturally occur in the environment and have virtually no background concentrations. As a consequence, these substances are ideal indicators of anthropogenic pollution. For instance, caffeine, ibuprofen and paracetamol can be used as indicators for contamination from untreated wastewater because of their high removal efficiency during waste water treatment (Warner et al., 2019). In contrast, the presence of chemicals that are generally poorly removed by waste water treatments, such as carbamazepine, may indicate contamination from treated as well as untreated waste water (Kahl et al., 2017). Tolyltriazole and hexamethoxymethylmelamine were suggested as suitable indicators of runoff water from roads (Seitz and Winzenbacher, 2017), while iodinated X-ray contrast media such as amidotrizoic acid, iothalamic acid, iomeprol and iopamidol were linked to wastewater from hospitals (Wolf et al., 2004). Some chemicals such as pesticides, personal care products (e.g., UV blockers), and pharmaceuticals (e.g., seasonal allergic reactions and infections) can be used to identify seasonal variation (Buttiglieri et al., 2009; Byer et al., 2011; Harman et al., 2011; Kasprzyk-Hordern and Baker, 2012a, 2012b; Loraine and Pettigrove, 2006). Combinations of multiple chemicals can be used in case indicators naturally occur in the environment. For instance, the ratio between caffeine and its metabolite paraxanthine can be used as an indicator of wastewater in areas where caffeine may occur naturally (e.g., cocoa and tea plantations) (Hillebrand et al., 2012a, 2012b). Polycyclic aromatic hydrocarbon (PAH) ratios have been used to distinguish between contamination resulting from direct residue of smelting activities associated with mining and its leachate (Warner et al., 2016), or between pyrogenic and petrogenic sources (Baragaño et al., 2022). Similarly, the ratio between metformin and guanyl urea can be used as an indication of untreated or poorly treated waste water (ter Laak et al., 2014).