Science of the Total Environment 933 (2024) 172827

Contents lists available at ScienceDirect

Science of the Total Environment

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

Apportioning sources of chemicals of emerging concern along an urban river with inverse modelling

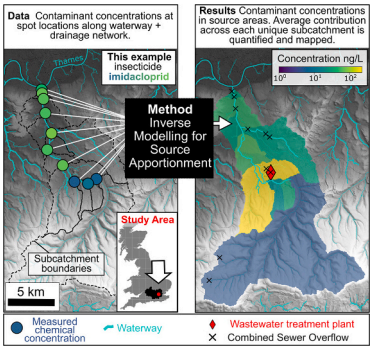
Kajetan Chrapkiewicz [a,*], Alex G. Lipp [b], Leon P. Barron [c], Richard Barnes [d], Gareth G. Roberts [a,*]

[a] Department of Earth Science and Engineering, Imperial College London, South Kensington Campus, London SW7 2AZ, UK
[b] Merton College, University of Oxford, Merton Street, Oxford OX1 4JD, Oxfordshire, UK
[c] MRC Centre for Environment and Health, Environment Research Group, School of Public Health, Imperial College London, Wood Lane, London W12 0BZ, UK
[d] Lawrence Berkeley National Laboratory, Wang Hall, Berkeley, CA 94720, USA

HIGHLIGHTS

- A new computational method of source apportionment of river contaminants is presented.
- It converts spot water samples into source concentration maps for individual chemicals.
- It uses the structure of waterway networks, is fast and quantifies uncertainty.
- Applied to an urban stream in London to link pesticides and drugs to diverse sources
- The method can be applied at scale with proposed optimal sampling Strategy.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Kevin V. Thomas

Keywords:
Pharmaceutical and illicit drugs
Contaminants of emerging concern
Water pollution
Inverse modelling
River mixing
Source apportionment
Thames drainage basin

ABSTRACT

Concentrations of chemicals in river water provide crucial information for assessing environmental exposure and risks from fertilisers, pesticides, heavy metals, illicit drugs, pathogens, pharmaceuticals, plastics and perfluorinated substances, among others. However, using concentrations measured along waterways (e.g., from grab samples) to identify sources of contaminants and understand their fate is complicated by mixing of chemicals downstream from diverse diffuse and point sources (e.g., agricultural runoff, wastewater treatment plants). To address this challenge, a novel inverse modelling approach is presented. Using waterway network topology, it quantifies locations and concentrations of contaminant sources upstream by inverting concentrations measured in water samples. It is computationally efficient and quantifies uncertainty. The approach is demonstrated for 13 contaminants of emerging concern (CECs) in an urban stream, the R. Wandle (London, UK). Mixing (the forward problem) was assumed to be conservative, and the location of sources and their concentrations were treated as unknowns to be identified. Calculated CEC source concentrations, which ranged from below detection limit (a few ng/L) up to 1 μg/L, were used to predict concentrations of chemicals downstream. Using this approach, >90% of data were predicted within observational uncertainty. Principal component analysis of calculated source concentrations revealed signatures of two distinct chemical sources. First, pharmaceuticals and insecticides were associated with a subcatchment containing a known point source of treated effluent from a

* Corresponding authors.
E-mail addresses: k.chrapkiewicz17@imperial.ac.uk (K. Chrapkiewicz), gareth.roberts@imperial.ac.uk (G.G. Roberts).

https://doi.org/10.1016/j.scitotenv.2024.172827
Received 15 January 2024; Received in revised form 25 April 2024; Accepted 25 April 2024
Available online 1 May 2024

wastewater treatment plant. Second, illicit drugs and salicylic acid were associated with multiple sources, interpreted as input from untreated sewage including Combined Sewer Overflows (CSOs), misconnections, runoff and direct disposal throughout the catchment. Finally, a simple algorithmic approach that incorporates network topology was developed to design sampling campaigns to improve resolution of source apportionment. Inverse modelling of contaminant measurements can provide objective means to apportion sources in waterways from spot samples in catchments on a large scale.

1. Introduction

Central to effective management of chemical contaminants in waterways (e.g., fertilisers, heavy metals, illicit drugs, pathogens, perfluorinated substances, pesticides, pharmaceuticals and plastics) is identifying their upstream sources and predicting their fate and effects downstream. However, rivers can efficiently mix the material they transport downstream, which provides both an opportunity and concomitant challenges for reliable source apportionment to support decisions for mitigation.

The opportunity arises because sampling water along waterways provides useful data to help characterise potentially large, complex and diverse environments upstream (e.g., Weltje, 1997; Bradley et al., 2017; Zhi et al., 2020; Lipp et al., 2020, 2021; Carraro et al., 2020). To take advantage of this opportunity, two challenges must be addressed. The first challenge is being able to measure a wide range of concentrations of large numbers of diverse contaminants, which may or may not have ecotoxicological relevance. Fortunately, new analytical methods have demonstrated that hundreds of such chemicals, including contaminants of emerging concern, can now be quantified rapidly down to the low-sub ng/L range in surface water using targeted direct-injection liquid chromatography-tandem mass spectrometry (LC-MS/MS) (Reemtsma et al., 2013; Campos-Mañas et al., 2017; Couchman et al., 2018; Egli et al., 2023). Direct analysis approaches using full-scan high-resolution accurate mass spectrometry (HRMS) can be used in a similar way to measure thousands of chemical residues using non-target and suspect screening, which can be further enhanced using machine learning tools (Richardson et al., 2021). Large databases of chemical properties, human and environmental toxicity, and exposure data also now exist, helping to prioritise risk assessment of chemicals (e.g., US EPA CompTox and the NORMAN Ecotoxicology Database). Recently, prediction of hazard directly from HRMS data has also been demonstrated (Feraud et al., 2023). It is therefore becoming more practical to identify, monitor and prioritise chemicals from thousands of samples annually.

The second challenge concerns how to interpret measured chemical concentrations along a waterway. From a theoretical perspective, the challenge is understanding how measured concentrations can be used to identify the source(s) of contaminants and their concentrations in source areas upstream. Multiple (known and unknown, point or diffuse) sources of contaminants that are mixed as they are transported downstream have to be (mathematically) disentangled from spot/grab sample measurements.

In this paper, inverse methodologies are used to objectively disentangle ('unmix' via computation) spot measurements of contaminant concentrations for source apportionment. A similar strategy has been used to establish sources of environmental-DNA in water samples and geologically relevant elements in fluvial sediments (Carraro et al., 2020; Lipp et al., 2020, 2021). As far as we are aware, this approach has not been used to identify sources of contaminants arising from urban consumption, which may have multiple influx routes. Once estimates of concentrations of contaminants in source areas exist, it is straightforward to estimate concentrations of contaminants throughout waterway networks via forward modelling, and to generate quantitative estimates of chemical fluxes and associated risks, potentially at the scale of entire drainage basins. Unlike existing models for predicting contaminant levels in waterways (e.g., Comber et al., 2013), this approach does not require any pre-existing estimates of upstream sources, requiring only observational data downstream.

This approach is demonstrated in a case study of the R. Wandle catchment, a tributary of the R. Thames in south London (Fig. 1; also see Fig. S1 of the Supporting Information). The Wandle is a polluted urban river (e.g., Egli et al., 2023). It is one of only a few hundred chalk streams globally that provide important habitats for various species of fish and invertebrates (e.g., Visser et al., 2019). Its network topology is nearly all natural, determined principally by topography (Fig. 1). In 2022, two monitored reaches of the R. Wandle, upstream at Carshalton and downstream of Croydon, were given 'poor' and 'moderate' ecological status, respectively.[1] Of the 16 proposed reasons for the downstream reach not achieving good ecological status, eight were related to pollution. Wastewater and urban runoff have been suggested by the Environment Agency (the environmental regulatory agency for England) as the principal sources of contaminants in the river. The catchment has diverse urban use including Beddington Farmlands, a landfill site. The nearby Beddington wastewater treatment plant (WWTP) discharges into the Wandle (Fig. 1c).

Egli et al. (2023) measured a suite of contaminant concentrations across London, including the R. Wandle in 2019–2021 (see Fig. 1 & S2). Pharmaceutical and illicit drug concentrations rose in the river such that they exceeded *predicted no-effect concentrations* (PNEC) reported in the NORMAN Network Ecotoxicology Database.[2] Other contaminants, including the insecticide imidacloprid, were observed at concentrations that constituted a high environmental risk (Perkins et al., 2024). We note that imidacloprid was banned in the EU for outdoor use, but has recently emerged again in the environment, potentially due to its use as a pet parasiticide. Major pathways to the environment probably include wastewater discharge and pet 'wash-off'. The source locations and overall contributions of the latter are especially challenging to determine.

The development of effective mitigation strategies would be aided if the exact sources of the contaminants could be identified. Whilst potential sources have been proposed, the exact location and magnitude of these sources remain unconstrained. The central question we address in this study is this: Can the sources of pollutants be identified objectively? This proof-of-concept study aims to develop approaches and recommendations that could be used for source apportionment elsewhere.

Existing mechanistic approaches to predict pollutant transport in waterways tend to assume that concentrations can be determined downstream by solving partial differential equations or via a mass-balance approach. Cox (2003) gives a succinct introduction to these approaches and to some of their limitations, which we expand upon below. Such approaches often predict the rate of change of concentrations by solving advection-diffusion equations, which can be expressed as $\partial C/\partial t = -v\nabla C + \kappa\nabla^2 C + S(x,t)$, where v sets the velocity at which contaminants advect downstream, κ is diffusivity, S is a source term, x is space (e.g., distance along a river) and t is time (e.g., Aster et al., 2005). Such partial differential equations can be modified to incorporate hydrological models (via the advective velocity and diffusivity terms), and transformation of chemicals downstream (via addition of terms representing in-stream biogeochemical processes). Pollutant sources,

[1] https://environment.data.gov.uk/catchment-planning/OperationalCatchment/3514.

[2] https://www.norman-network.com/nds/ecotox.

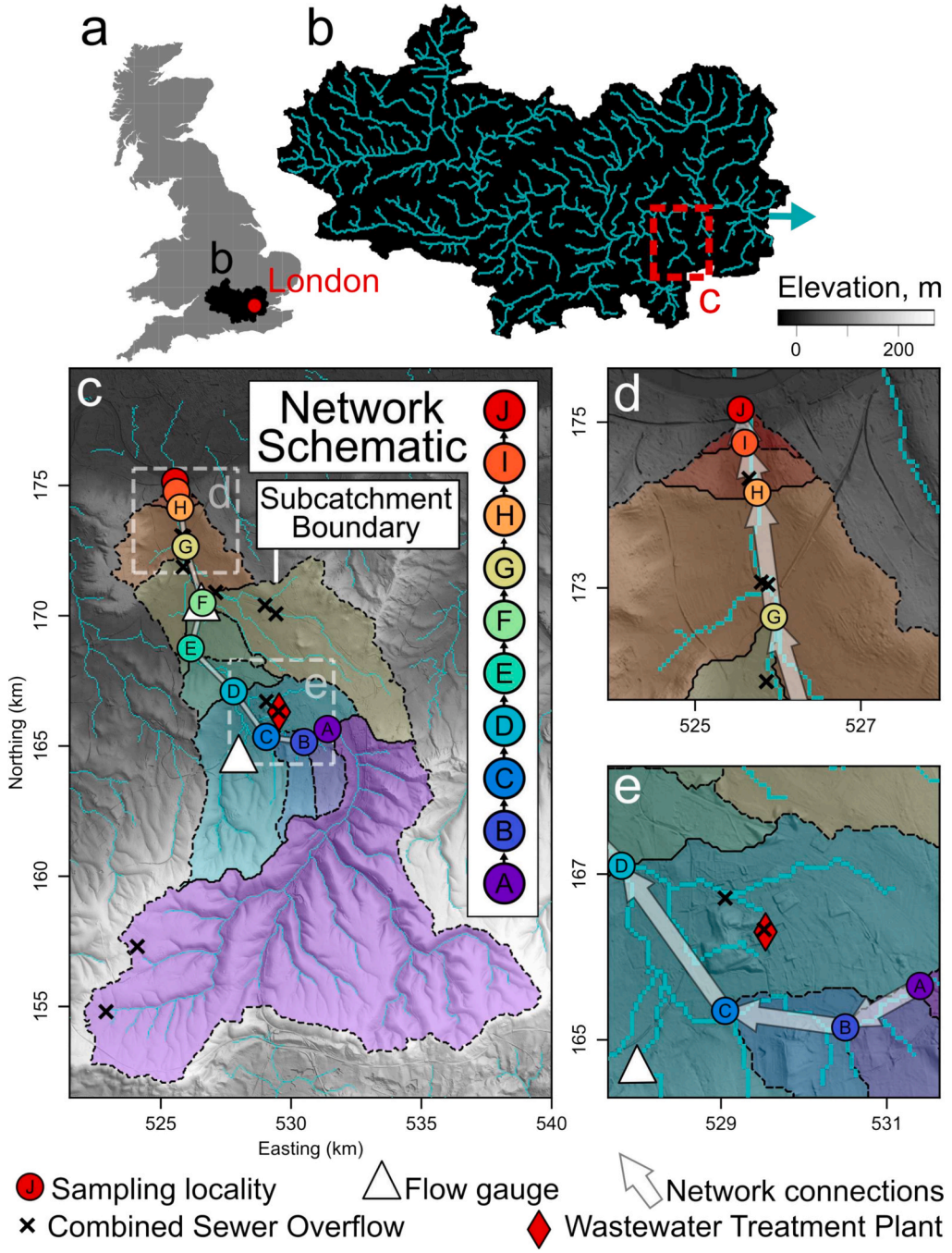


Fig. 1. Study area. (a–b) Study area (red dashed polygon) with respect to the UK, London and Thames basin (b); rivers and basin outline are from HydroSHEDS (Linke et al., 2019). Arrow = Thames Estuary. (c) Wandle catchment. Hill-shaded topography from 10-m resolution LIDAR digital elevation model. Turquoise lines = rivers with upstream area > 1 km^2. Circles = colour-coded 10 localities where water samples were collected and used to predict composition of upstream source regions. Coloured regions = unique upstream subcatchments of each locality; these colours are used consistently throughout the paper. White triangles = flow gauges, the one closest to locality F was used to calculate flux. x = CSOs that were active at any time during the sample collection period. Red diamond = Beddington wastewater treatment plant. Inset shows network schematic. (d–e) Zoom into most densely sampled subcatchments.

generated from land-use models or arbitrarily, say, can be inserted into such models (e.g., in the SAGIS decision support tool; Comber et al., 2013). Such approaches are widely used but they tend to require a priori information that is not always available and can be highly uncertain (e.g., concentrations of contaminants in source areas, relationships between land-use and pollutant export, contaminant travel times, hydrological conditions). In this study, we explore whether reliable source apportionment can be achieved without recourse to such models, and without the need for poorly constrained parameters.

A widely used approach to source apportionment makes use of multivariate and/or other statistical techniques (e.g., Weltje, 1997; Zhou et al., 2023; Chen et al., 2023). Such studies tend to seek 'theoretical' source compositions that explain the data, or instead by clustering measured concentrations into categories. These statistical methods can qualitatively identify discrete types of source, but struggle to quantify their importance. Alternatively, a suite of complimentary approaches exists that seek to adjust concentrations or fluxes of a priori assumed sources (end-members) to fit observations (e.g., Christophersen et al., 1990; Collins et al., 2020) Such approaches are therefore limited in cases where source compositions are variable or not well characterised, or

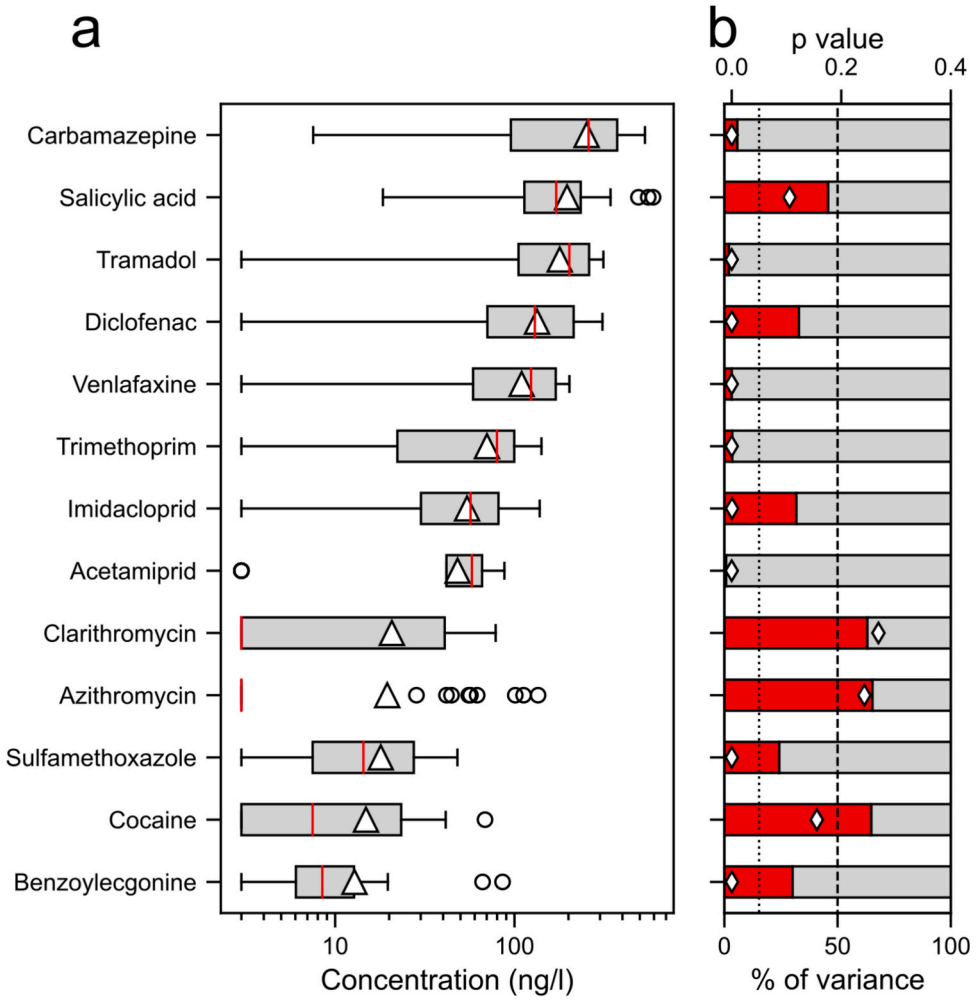


Fig. 2. Analysis of variance of chemicals in Wandle catchment. (a) Concentration of chemicals at localities A–J including all replicates (33 measurements per chemical, see Fig. 1). Note that the limit of detection (LOD) = 3 ng/L. Triangle = mean. Red vertical line = median. Box encloses 25th to 75th percentiles. Whiskers = ±1.5× the inter-quartile range. Circles = outliers, i.e., greater or smaller than values indicated by whiskers. Note: median for azithromycin equals LOD, as it was not detected in more than half of the samples. (b) Analysis of Variance (ANOVA). *p*-values = white diamonds. Dotted line = *p*-value of 0.05. Red/grey bars = intra−/inter-site variances normalised to 100 %. Dashed line (50 %) indicates equal intra-site and inter-site variance.

there exist unknown contaminant sources within a study area.

In contrast, inverse modelling using network topology has been demonstrated as a useful approach for directly and objectively apportioning chemical concentration measurements to actual upstream sources (e.g., Milledge et al., 2012; Carraro et al., 2021; Lipp et al., 2021; Barnes and Lipp, 2024). Such approaches have the benefit of not needing to assume composition, concentrations or locations of sources *a priori*.

In this work, inverse modelling of chemical concentrations measured at spot locations along a waterway is used to directly solve for actual locations and concentrations of contaminants in source areas upstream. The approach uses the topology of waterway networks, a simple theory of contaminant mixing downstream (the forward model) and optimisation theory. The inverse model seeks sources (their locations and concentrations) that, when mixed by the forward model, predict concentrations downstream that best fit observed (measured) chemical concentrations (Barnes and Lipp, 2024).

The structure of this paper is as follows. Data and methodologies are first described. The inversion strategy for recovering the sources of thirteen chemicals in the Wandle catchment is then demonstrated. Principal component analysis is then applied to the calculated sources to identify sources of contaminants with similar properties. Once armed with source concentrations, chemical exposure throughout the drainage basin is calculated. The results from inverse modelling are used to calculate contaminant fluxes and risk quotients. The results emphasise the need to consider the topology of waterway networks when choosing sampling locations. Thus, in the second part of this paper, an objective approach to designing sampling campaigns, optimised to provide unbiased identification of contaminant provenance and quantify associated fluxes of chemicals downstream, is developed.

2. Data & methods

2.1. Selection of chemical data for analysis

The chemical data used in this study are described in detail by Egli et al. (2023). In summary, 390 samples were collected from waterways across Greater London in 2019–2021. Using novel direct-injection LC-MS/MS, 10,029 concentration measurements of 66 unique contaminants of emerging concern were generated including pharmaceuticals, pesticides and illicit drugs, which, alongside rainfall, river flow rate, and CSO event duration monitoring data, provide contextual information for this study (see Fig. S3–6; see also Munro et al., 2019). This study is focused on the R. Wandle catchment where $N = 33$ samples were collected during two campaigns in two consecutive years (2020–2021; Figs. 1 & 2). Each campaign spanned approximately two months in autumn or winter. All sampling localities were upstream of the R. Thames tidal

reach. Summary statistics for the 66 chemicals measured along the R. Wandle are shown in Fig. S3. For all compounds in the analytical method, the mean and standard deviation of the estimated limit of detection (LOD) and lower limit of quantitation (LLOQ) were 3 (± 5) ng/L and 9 (± 17) ng/L, respectively (Egli et al., 2021).

For the remainder of this paper, we focus on investigating thirteen contaminants of emerging concern (CECs) in the Wandle (Fig. 2a & S7-S19). They included those contaminants with (a) the highest and consistently measurable concentrations (see Fig. S3), (b) potential for environmental risk (including several from EU Water Framework Directive 'Watch-Lists'), and diverse (c) physicochemical properties (logP, pKa, logBCF, $\log K_{oc}$, molecular weight, etc.) and (d) PNEC values (see Table S1). For the latter two criteria, a wide range was required so that results from inverting chemical concentration measurement data could be more reliably assessed. These chemicals provide a diverse suite of tracers for investigating multiple sources, pathways and fates of contaminants in waterway networks. The contaminants investigated were carbamazepine (anticonvulsant), salicylic acid (a metabolite of aspirin, also used as an active ingredient in skincare products), tramadol (opioid analgesic), diclofenac (anti-inflammatory), venlafaxine (antidepressant), imidacloprid and acetamiprid (insecticides), cocaine and its metabolite benzoylecgonine, and trimethoprim, clarithromycin, azithromycin and sulfamethoxazole (all antibiotics). Concentrations for all chemicals studied, including repeat measurements and frequency of detection (33 measurements of each chemical), at individual localities (replicates), are shown in Fig. 2a and S7e–19e (see Egli et al., 2023, for original dataset).

2.2. Forward (mixing) model assumptions

For simplicity, this exploratory study assumes that concentrations of contaminants at each sample site do not vary as a function of time. Pollutants are assumed to be well mixed across the channel (i.e., constant at a given streamwise distance) and conservatively mixed downstream (of unknown sources; see Eq. 8). Measured chemical concentrations are assumed to be representative of chronic contamination. This forward modelling strategy assumes that chemical concentrations downstream are at a steady state and well mixed. For simplicity and to test the utility of a simple mixing model for describing the observations, it purposefully does not incorporate travel times of chemicals along rivers, nor decay or other chemical transformation.

2.3. Variance of measured concentrations

To explore whether measured concentrations, c, from the Wandle catchment accord with the assumptions described in Section 2.2, a nested analysis of variance (ANOVA) for all sample concentrations was performed. Each contaminant was assessed independently. Inter- (i.e., between localities) and intra-site (i.e., within locality) variance was calculated for the thirteen contaminants measured at the ten localities sampled in 2020–2021 (Fig. 2b). The inter-site variance reflects the systematic spatial variability of contaminant concentrations introduced by, say, different sources distributed across the catchment. The intra-site variance reflects the variability of contaminant concentration at the same site in time, introduced by, say, inefficient mixing of the contaminant or temporally variable sources.

Samples collected at different dates, but located within 100 m of each other, were considered to be replicates and grouped into localities A–J using the DBSCAN clustering algorithm in the sklearn.cluster Python library (Fig. 1; see also Fig. S7–S19). The results of the clustering were then quality-checked by visually inspecting the location of the clusters on a map. The following analysis of variance incorporates all samples at the $n = 10$ (A–J) localities, where the ith locality contains n_i samples. As concentration data was log-normal, log-variance was explored. Variation of contaminant concentration within each locality (intra-site variation) is

$$v_w^i = \sum_{j=1}^{n_i} \left(\log c_j - \log \bar{c}_i\right)^2, \tag{1}$$

where j is the sample index unique to each locality, c_j is the concentration of individual samples within the locality and $\bar{c}_i$ is the mean concentration of all samples at the locality,

$$\bar{c}_i = exp\left(\frac{1}{n_i}\sum_{j=1}^{n_i} \log c_j\right) \tag{2}$$

and is equal to the geometric mean of the concentrations $\left(\prod_{j=1}^{n_i} c_j\right)^{1/n_i}$. Similarly, the standard deviation at ith locality is defined as

$$\sigma_{c_i} = exp\sqrt{\frac{1}{n_i}\sum_{j=1}^{n_i}\left(\log c_j - \log \bar{c}_i\right)^2}. \tag{3}$$

Variation incorporating all samples at all localities is

$$v_w = \sum_{i=1}^{n} v_w^i = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(\log c_j - \log \bar{c}_i\right)^2. \tag{4}$$

Variation between localities (inter-site variation) is,

$$v_b = \sum_{i=1}^{n}\left(\log \bar{c}_i - \log \bar{c}\right)^2. \tag{5}$$

where $\bar{c}$ is the grand mean of all samples at all localities

$$\bar{c} = exp\left(\frac{1}{N}\sum_{j=1}^{N} \log c_j\right). \tag{6}$$

The ratio of (arithmetic) mean variation within a locality (intra-site) and between localities (inter-site) can thus be expressed as

$$\frac{\sum_{i=1}^{n} v_w^i}{n \cdot v_b}, \tag{7}$$

and is shown on Fig. 2b. The length of the red and grey bars in this figure show within-locality (intra-site) and between-localities (inter-site) variance, respectively. The dashed line on this panel indicates equal variance. If calculated p-values, indicated by the white diamonds, are $>$ 0.05 (dotted line) the null hypothesis—the difference in intra- and inter-site variance is not statistically significant—cannot be rejected.

In the source apportionment calculations that follow, the mean location of replicates at each locality A-J, their geometric mean concentration and the corresponding standard deviation (Eqs. 2 & 3, respectively) are used.

2.4. Drainage network and sample localities

A digital drainage network for the study area was generated using established flow routing techniques as follows (Fig. 1). First, flow directions of waterways were calculated with LandLab package (Hobley et al., 2017) by applying the D8 algorithm (O'Callaghan and Mark, 1984) to the Center for Ecology and Hydrology's Integrated Hydrological Digital Terrain Model, which has a horizontal resolution ~ 50 m (Morris and Flavin, 1994). Overland flow was defined for grid cells with upstream drainage areas >1 km^2. Calculated drainage compares favourably to independently mapped waterways (Ordnance Survey, Environment Agency) and high-resolution (10-m) LIDAR topographic data. Localities were then automatically snapped to the nearest node on the digital drainage network. The resultant map was manually checked for mislocations. Only locality F required a small manual relocation from an adjacent tributary onto the main channel.

2.5. The forward model: predicting chemical concentrations downstream from known sources

A simple approach, using only the topology of a drainage network, was used to solve the forward problem, i.e., predict pollutant concentrations downstream given a set of sources upstream (see Barnes and Lipp, 2024). Concentration d at a given point downstream was calculated based on contaminant concentrations c_i and fluxes q_i (kg/s) of its N upstream sources such that

$$d = \frac{1}{Q}\sum_{i=1}^{N} q_i c_i \quad \text{where} \quad Q = \sum_{i=1}^{N} q_i. \tag{8}$$

Q is the total flux of the river. Note that the units of d and c_i must be the same (e.g., ng/L, kg/m^3), but are otherwise arbitrary (i.e., they could be any measure of concentration, e.g., ppm, ppb, %). This simple forward model can predict concentrations for all points along the drainage network. However, measurements used to validate predicted concentrations are typically available only at select locations (e.g., Fig. 1). Consequently, it is often adequate to calculate d only at the locations where the measurements are available and thus reduce the computational costs.

The approach introduced in Barnes and Lipp (2024) was used to predict contaminant concentrations $d_1, d_2, \ldots, d_N$ at N localities $x_1, x_2, \ldots, x_N$ of spot samples along the drainage network. The method uses a graph representation of the nested samples based on their connectivity within the drainage network. A topological network (a directed acyclic graph) of unique subcatchments is generated from the sample-set (see e.g., inset in Fig. 1c). Each node on the network corresponds to a subcatchment defined by a sample site. Each subcatchment, which encloses the drainage area between adjacent samples, is unique i.e., does not overlap with any other subcatchment. Each subcatchment (node) is described by three parameters: unique upstream area A (m^2), run-off ϕ $(\mathrm{kg/s/m}^2)$, here defined as a mass flux per unit upstream area, and concentration $(\mathrm{kg/m}^3)$. For each subcatchment, average run-off can be combined with its drainage area to calculate the flux, $q_i = A_i\phi_i$. In this study, run-offs are assumed identical for all the subcatchments so, effectively, the fluxes depend only on the subcatchment areas. We stress that subcatchments in this study were determined by the distribution of sample sites and the topology of the waterway network.

For each (ith) locality, a set $\mathbb{U}_i$ of all its upstream subcatchments is defined. For example, for subcatchment D the set $\mathbb{U}_i = \{\mathrm{D, C, B, A}\}$. The forward operator can then be expressed in discrete form as

$$d_i = \frac{1}{\sum_{j\in\mathbb{U}_i} q_j}\sum_{j\in\mathbb{U}_i} c_j q_j. \tag{9}$$

When upstream basins are nested, the forward model (Eq. 9), is solved in topological order, building the above expression for each node upstream to downstream (see inset in Fig. 1c).

2.6. Inverse method

In this study, sparse downstream samples were inverted to determine source concentrations upstream. This approach sought to identify subcatchment concentrations that predict (via Eq. 9) concentrations downstream that minimise misfit to spot measurements (Fig. 1). We emphasise that this approach seeks to quantify optimal concentrations of contaminants in source areas (subcatchments) upstream, rather than theoretical end-members. It seeks solutions without assuming any a priori knowledge of source concentrations. The inverse method introduced by Barnes and Lipp (2024), which uses the graph-based forward model described above, was used.

More formally, this approach was used to identify the vector of source concentrations, $\mathbf{c}$ $(= c_1, c_2, \ldots, c_n)$, that best fits the vector of measured chemical concentrations downstream $\mathbf{d}^{\mathrm{obs}}$ $\left(= d_1^{\mathrm{obs}}, d_2^{\mathrm{obs}}, \ldots, d_n^{\mathrm{obs}}\right)$. The inverse model aimed to minimise the cumulative misfit between $\mathbf{d}^{\mathrm{obs}}$ and the vector of predicted concentrations at the same localities, $\mathbf{d}^{\mathrm{pred}}$ $\left(= d_1^{\mathrm{pred}}, d_2^{\mathrm{pred}}, \ldots, d_n^{\mathrm{pred}}\right)$, calculated using the forward model (Eq. 9). Formally,

$$min\sum_{i\in N} \text{misfit}\left(\mathbf{d}^{\mathrm{obs}}, \mathbf{d}^{\mathrm{pred}}\right), \tag{10}$$

was sought. Since concentration data can span many orders of magnitude, and because efficient convex optimisation was desirable, the following misfit function was defined

$$\text{misfit}\left(\mathbf{d}^{\mathrm{obs}}, \mathbf{d}^{\mathrm{pred}}\right) = max\left(\frac{\mathbf{d}^{\mathrm{obs}}}{\mathbf{d}^{\mathrm{pred}}}, \frac{\mathbf{d}^{\mathrm{pred}}}{\mathbf{d}^{\mathrm{obs}}}\right). \tag{11}$$

2.6.1. Regularisation

Measurements are likely to contain noise and analytical uncertainties and the forward model is a simplification of reality. Thus regularised ('simple') solutions for each contaminant are sought. Deviations of $\mathbf{c}$ from the (geometric) mean of observed concentrations (at the sample sites), $\overline{\mathbf{d}}$, are penalised, where

$$\overline{\mathbf{d}} = \left(\prod_{i=1}^{n} \mathbf{d}^{\mathrm{obs}}\right)^{1/n}. \tag{12}$$

Thus, the optimisation algorithm works to solve the following

$$\min_{\text{w.r.t. c}} \left\{\sum_{i\in N} max\left(\frac{\mathbf{d}^{\mathrm{obs}}}{\mathbf{d}^{\mathrm{pred}}}, \frac{\mathbf{d}^{\mathrm{pred}}}{\mathbf{d}^{\mathrm{obs}}}\right) + \lambda\sum_{i\in N} max\left(\frac{\overline{\mathbf{d}}}{\mathbf{c}}, \frac{\mathbf{c}}{\overline{\mathbf{d}}}\right)\right\}, \tag{13}$$

by seeking optimal $\mathbf{c}$. Regularisation was controlled by the value of λ. The optimal model was the one that minimised the terms within the braces in Eq. 13. The optimal values of λ were identified for each chemical via a systematic sweep of values in a series of inverse models (see e.g., Fig. S20-32c). Each inverse model solved for source concentrations by minimising misfit to the mean values of measured concentrations at localities A–J. Optimal values for each chemical were identified as those that yielded maximum curvature when data misfit (first term in braces in Eq. 13) was compared to model misfit (second term in braces in Eq. 13; see Parker, 1994). Solutions to Eq. 13 were sought using *embedded conic solvers* within the cvxpy Python package (Domahidi et al., 2013; O'Donoghue et al., 2016).

2.6.2. Uncertainty propagation

The impact of measurement uncertainties and noise (e.g., time-dependent changes in concentrations at individual sample sites) on predicted concentrations were assessed via Monte Carlo experimentation as follows. A log-normal distribution of uncertainties, unique to each sample site and chemical, with mean and standard deviation defined by Eq. 2 & 3 was assumed for each contaminant from which $n = 10^4$ 'data' vectors were drawn randomly, effectively adding noise to measured concentrations. Regularisation parameters (per contaminant) were held constant in all iterations of the Monte Carlo experiment, so that the mean, median, and standard deviation of the ensemble of final models were straightforward to interpret. Median values were calculated to assess the skewness of the distribution (e.g., Shapiro and Ritzwoller, 2002).

2.7. Environmental risk assessment

For each chemical, s, and each subcatchment, i, risk quotients were calculated such that $R_i^s = c_i^s/p_s$, where p_s indicates PNEC values from the NORMAN ecotoxicological database, with the following values (ng/L) acetamiprid: 24, azithromycin: 19, carbamazepine: 2000, cocaine: 2456, diclofenac: 50, imidacloprid: 13, salicylic acid: 18,000, sulfa-

methoxazole: 600, tramadol: 8653, trimethoprim: 120,000, venlafaxine: 880. Following Egli et al. (2023) and Palma et al. (2014), risks were defined as *high* if $R > 10$, *medium* if $10 \geq R \geq 1$, *low* if $1 \geq R \geq 0.1$, and *insignificant* if $R < 0.1$. See Fig. 4 for calculated concentrations, c_i, and uncertainties, σ_c, used to calculate R and uncertainties shown in Fig. S33.

2.8. Estimating chemical fluxes downstream

With estimates of both source concentrations and river discharge, chemical fluxes were calculated, q (kg/s), for each subcatchment. Here, the mass flux of a chemical from a subcatchment was estimated using the product of water discharge, f $\left(\mathrm{m}^3/\mathrm{s}\right)$, and volumetric density (concentration) of the chemical, c $\left(\mathrm{kg}/\mathrm{m}^3\right)$,

$$q = f \cdot c. \tag{14}$$

Uncertainties in concentrations and water discharge were propagated to calculate uncertainties in chemical flux,

$$\sigma_q = \sqrt{\left(\frac{\partial q}{\partial c}\sigma_c\right)^2 + \left(\frac{\partial q}{\partial f}\sigma_f\right)^2} = \sqrt{\left(f\sigma_c\right)^2 + \left(c\sigma_f\right)^2}, \tag{15}$$

where c and σ_c are the respective mean and standard deviation of concentrations extracted from the ensemble of models generated by Monte Carlo experimentation. Here, f and σ_f are the mean and standard deviation of water discharge. Errors were assumed to be independent (Taylor, 1997). The values were not corrected for smoothing due to regularisation.

Water discharge was estimated during the sampling campaign using measurements made between 2020 and 2021 at the "Wandle at South Wimbledon" gauge station (UK-CEH's National River Flow Archive: ID: 39003; Fig. S6). Calculated mean discharge and standard deviation are $f = 1.84$ and $\sigma_f = 0.15$ m^3/s, respectively. Average unit $\left(\text{per } \mathrm{m}^2\right)$ runoff, $\mathscr{O}\left(10^{-8}\ \mathrm{m}/\mathrm{s}\right)$, was estimated by normalising measured discharge by the drainage area upstream of the gauge station $\left(176\ \mathrm{km}^2\right)$. Note that this measurement, which was extracted from the digitised drainage network, substantially exceeded the effective drainage area estimated by UK-CEH $\left(54\ \mathrm{km}^2\right)$. The consequences of changing the run-off were also assessed. Finally, water discharge at the localities A–J was calculated by multiplying average runoff by the unique catchment area upstream of each locality.

2.9. Principal component analysis of source concentrations

Whilst exploring the spatial distribution of sources of each chemical individually is important, it may also be useful to identify common patterns between multiple chemicals. Such patterns may, for instance, reflect a smaller number of underlying contaminating processes or sources manifesting in measurable concentrations. Identifying such patterns may facilitate more efficient mitigation. To assess similarities and dissimilarities in the variance of calculated concentrations, principal component analysis (PCA) was used to find eigenvectors of the covariance matrix of calculated subcatchment concentrations of chemicals. The inverse modelling results provided concentrations of n chemicals in m subcatchments. For stability, PCA requires $m > n$, so four chemicals were arbitrarily excluded. To assess the sensitivity of PCA results to the choice of removed chemicals we performed *bootstrapping* using all 715 combinations of removed chemicals. To avoid *closure effects*, concentrations were recast using the centred log-ratio (clr) transformation prior to PCA (Aitchinson, 1983).

2.10. Identifying 'optimal' sampling sites

In general, sources of contaminants in waterways are unknown and resources are finite. Thus, it is desirable to identify where in a catchment samples should be taken such that recovered sources are most informative. A variety of methods for objective (unbiased) identification of sampling sites on drainage networks have previously been proposed which seek to optimise sampling for different purposes (e.g., Dixon et al., 1999; Carraro et al., 2021; Singhal et al., 2023). For our specific purposes, it would be useful to identify sampling localities that permit source apportionment to n equal area subcatchments. The approach we used is designed to provide unbiased coverage in a catchment where contaminant sources could be anywhere with equal probability. First, a target catchment size, A, is defined. This value determines sample localities required to divide the network into subcatchments no greater in area than A. The scheme developed here was parameterised using topography extracted from a digital elevation model so that optimal sampling localities in a natural drainage network could be identified. The algorithm is provided and described in detail here: zenodo.org/records/7311352.

3. Results

3.1. Analysis of variance

The analysis of variance (Fig. 2) indicates that the concentrations of most chemicals, such as imidacloprid and benzoylecgonine were controlled by chronic, temporally invariant, pollutant sources distributed inhomogeneously across the catchment. In contrast, concentrations of chemicals such as clarithromycin and azithromycin may have been controlled by local effects or temporally variable sources. These results suggest that for most contaminants, inverting their concentrations measured at spot localities along a river should yield useful information about the location of their sources and associated concentrations.

3.2. Inverse modelling for source concentrations

3.2.1. Validating the method by inverting synthetic data

The capability of the inverse model to apportion chemicals to sources upstream was tested by first inverting synthetic data (see Fig. S34). In a series of experiments, known concentrations of an arbitrary contaminant were assigned to subcatchments and the regularisation parameter was varied systematically such that $\lambda = 10^{-2}$, 10^{-4} or 10^{-6}. In the simple scenarios examined, the 'true' concentration of the contaminant in single subcatchments was set to be arbitrarily high (1000 ng/L), and arbitrarily low (1 ng/L) in all other subcatchments. For instance, setting the concentration in subcatchment D to be very high compared to other subcatchments could represent a simplified example of pollution from a single point-source in subcatchment D, for example, the Beddington wastewater treatment plant. These known 'source' concentrations were used to calculate concentrations at localities A–J downstream using the forward model (Eq. 9).

Calculated concentrations at localities A–J were then inverted to recover source concentrations (Eq. 13). For clarity, the only inputs into the inverse models were the drainage network, the synthetic concentrations of the contaminant at the ten localities, and the value of the regularisation parameter, λ. In other words, the original 'known' *source* concentrations were discarded. Instead, they were recovered by inverting the concentration 'measurements' at localities A–J.

When $\lambda = 0$ (i.e., no regularisation) predicted chemical concentrations at sources matched the 'known' true concentrations accurately (within machine precision). When λ was increased, calculated concentrations more closely aligned with the average concentration of the ten localities, as expected. Peak concentrations were always correctly located in subcatchment D for all values of λ tested. These results indicate that source apportionment from inversion of real chemical observations along the R. Wandle could identify source locations and quantify contaminant concentrations there.

An important concern was that noise may have negatively impacted

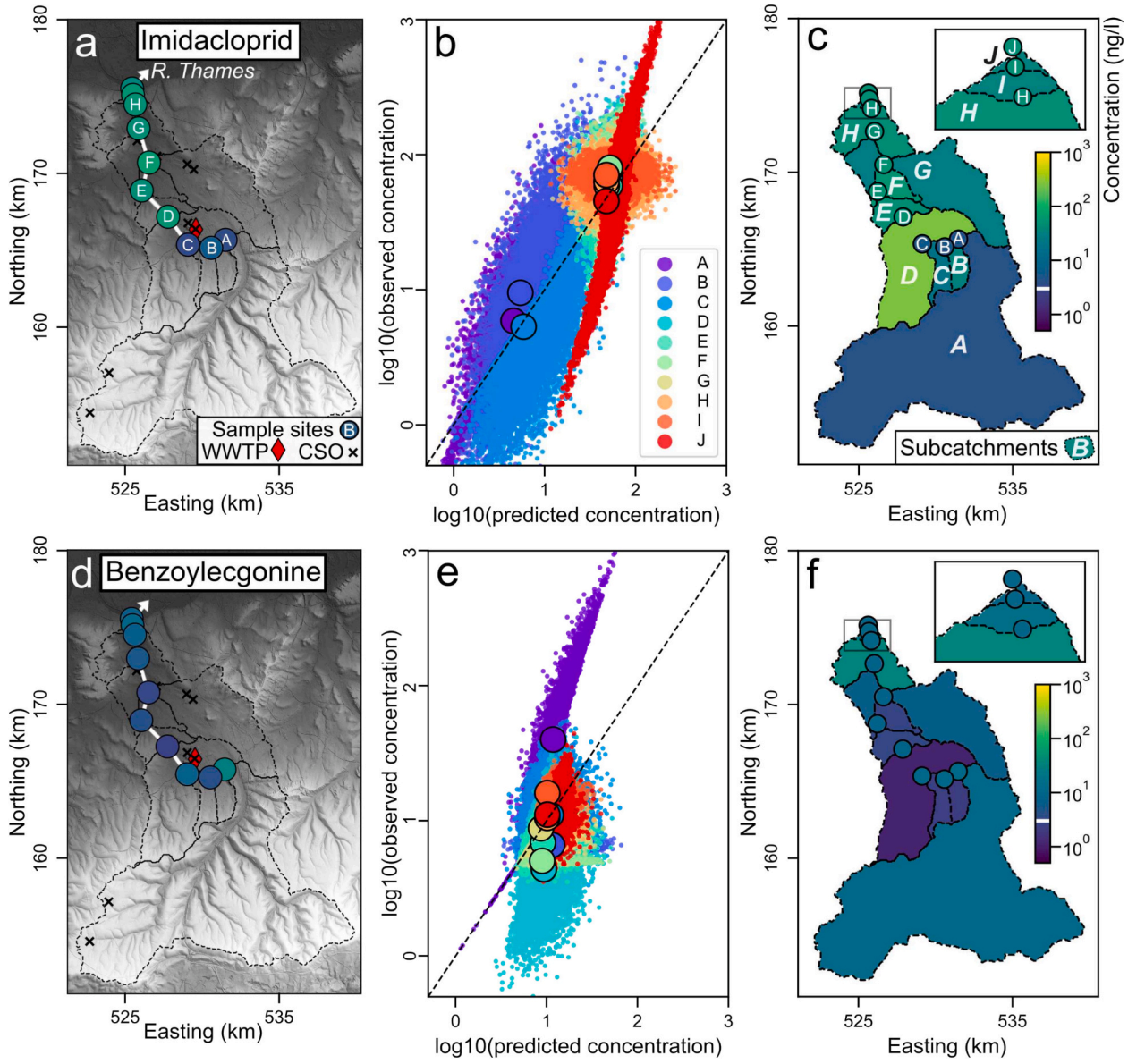


Fig. 3. Objective source apportionment for insecticide imidacloprid and cocaine's metabolite benzoylecgonine. (a) Observed (measured) imidacloprid concentrations along the R. Wandle (coloured circles); thick white line = sample network; thin dashed curves = subcatchments unique to each sampling locality (A–J); grey scale = hill-shaded digital elevation model (see Fig. 1 and key for additional annotations). (b) Observed vs. predicted downstream concentrations at sample localities A–J (small circles; see Fig. 1) from Monte Carlo experiment incorporating 10^4 inverse models of measured concentrations randomly perturbed within their uncertainties unique to each observation; large circles = mean misfit for each locality. Dashed line = 1:1 correlation. Downstream predictions were calculated from the results of apportionment (i.e., predicted upstream concentrations; panel c) by solving the forward problem. (c) Mean predicted upstream concentrations in each subcatchment for optimal regularisation. Coloured circles = mean observed concentrations at sample localities (panel a). White line on colour bar = detection limit. (d–f) Results for benzoylecgonine. See Fig. S22 & S27 for extended results. Fig. 4 summarises results for all the chemicals inverted in this study.

the recoverability of source concentrations from measurements downstream. Noise might be introduced by analytical error or as a consequence of physical or chemical processes not incorporated into the forward model (e.g., rivers that are not well-mixed across a channel), for instance. Fig. S35 shows the results from 10^4 inversions of synthetic data generated by the procedure described in the previous paragraph plus added random noise. In each of these experiments, noise was added to the (initially noise-free) mean chemical concentrations at localities A–J shown in Fig. S34. Noise was added so that the range of values at each locality encompassed the range of imidacloprid field data, an analyte that could have multiple point and diffuse sources including, for instance, dogs swimming in the river and WWTP effluent. The results indicate that inverting spot measurements of concentrations (even with added noise) could apportion contaminants to sources reliably.

Results from varying the location of the source of the arbitrary contaminants are shown in Fig. S36–37 (see Supporting Information for details). The *resolution matrices*, quantitative measures of the inverse model's ability to resolve model parameters, shown in these figures, demonstrated that inverse modelling of real chemical data could be expected to recover the locations and concentrations of contaminants in source areas. They also demonstrate that increasing noise and regularisation 'smears' calculated concentrations between nearby subcatchments.

3.2.2. *Inverting measured chemical concentrations*

Fig. 3a–c shows the results from inverting concentrations of the pesticide imidacloprid measured at localities A–J (Fig. 1). In these models, regularisation parameter $\lambda = 10^{-1.7}$, which sits at the location

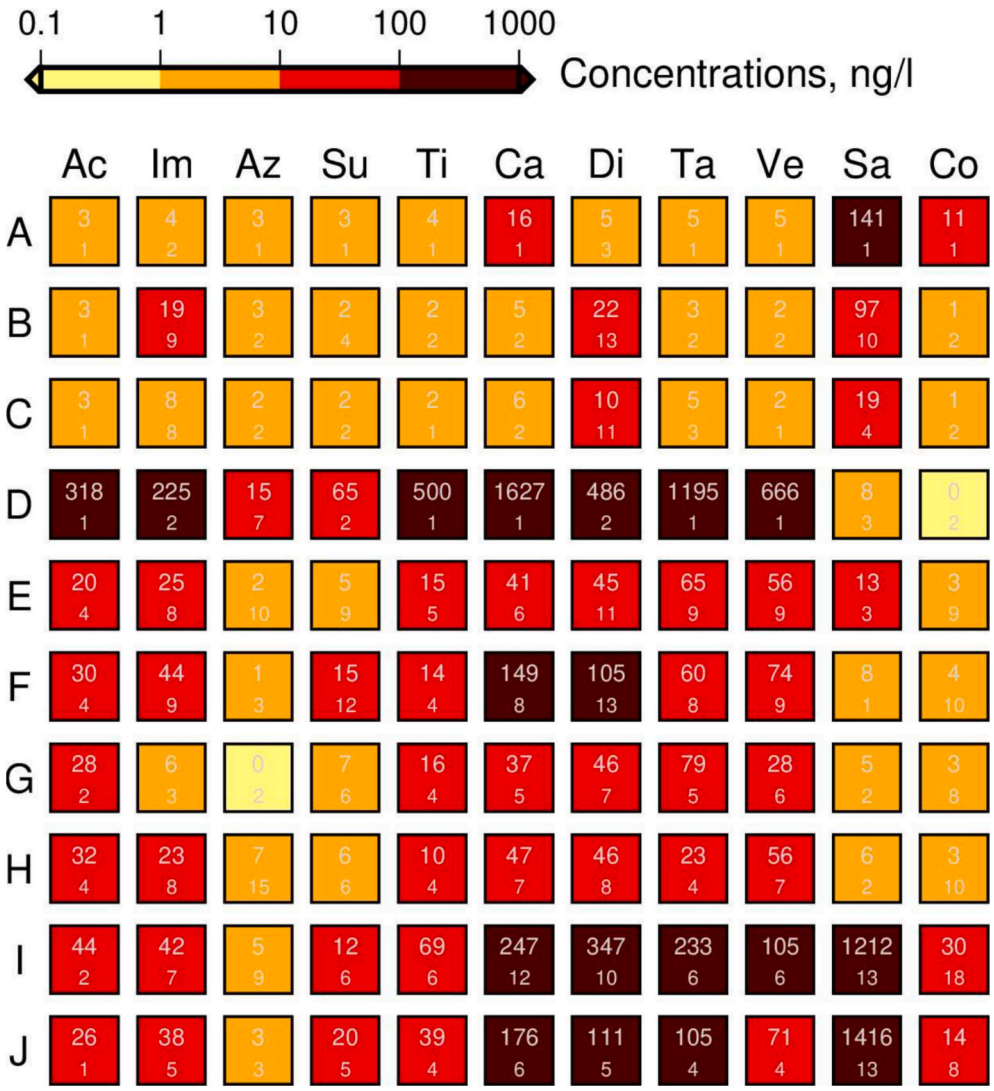


Fig. 4. Summary of source apportionment of contaminants to subcatchments A–J. Calculated mean concentrations and standard deviations (ng/L) from Monte Carlo inversion (see Fig. S20–S32 for extended results); abbreviations—Ac: acetamiprid (24), Im: imidacloprid (13), Az: azithromycin (19), Cl: clarithromycin (−), Su: sulfamethoxazole (600), Ti: trimethoprim (120,000), Ca: carbamazepine (2000), Di: diclofenac (50), Ta: tramadol (8653), Ve: venlafaxine (880), Sa: salicylic acid (18,000), Be: benzoylecgonine (−), Co: cocaine (2456); numbers in parentheses indicate *predicted no-effect concentration* (PNEC; ng/L; NORMAN Ecotoxicology Database; see Discussion). Note that for very low concentrations calculated standard deviations may suggest negative concentrations but they are in fact lower-bounded by 0. See Fig. S33 in Supporting Information for associated risk quotients.

of maximum curvature in the plot of data misfit vs. model roughness (see Sect. 2.6; Fig. S20-32c). Fig. 3b shows a comparison of observed and predicted concentrations at the sample sites for all 10^4 iterations of the Monte Carlo experiment, mean values for each subcatchment are also shown. These results indicate that the inverse model generated good fits to the data. The resultant source apportionment is shown in panel (c). Alternative presentation of fits to data, source apportionment and uncertainties are shown in Supporting Information.

The relatively high calculated concentrations in subcatchment D emphasise its importance as a source of imidacloprid. This compound has recently emerged as of concern following a market repositioning to increased use as a pet parasiticide. It is poorly removed during wastewater treatment and, as a result, is regularly present in receiving waters impacted by WWTP effluents (Perkins et al., 2024). In D the median source concentration across all models is more than an order of magnitude higher than in all other subcatchments. Inverse modelling of synthetic data discussed in the previous section (i.e., a model with a single subcatchment with high concentrations) reveals a very similar spatial pattern of recovered concentrations with artificially increased values in subcatchments B–C and E–F for high values of λ (see Fig. S34–35). As a result, a portion of the relatively higher calculated concentrations for adjacent subcatchments, i.e., B–C and E–H could be attributed to regularisation. The results are insensitive to the LOD (or LLOQ) values tested (Fig. S39–41).

Inverse modelling of benzoylecgonine, a metabolite of cocaine, is shown in Fig. 3d–f. The same modelling strategy used to invert imidacloprid was followed (with $\lambda = 10^{-2.7}$; Fig. S22c). In contrast with imidacloprid, calculated concentrations of benzoylecgonine in mid-reach subcatchments (B–F) of the Wandle were low. We note that cocaine and benzoylecgonine are generally very well removed in wastewater treatment (Munro et al., 2019). Calculated concentrations were highest in subcatchments upstream (A) and downstream (G–J), which did not include WWTPs. Predicted concentrations of benzoylecgonine are the most poorly fitting of all the chemicals inverted. Residual misfit between observed and theoretical concentrations of benzoylecgonine is highest at localities A and D (see Fig. S22a), which is similar to the results obtained for cocaine. Calculated concentrations at other localities match observations much more closely. In general,

> 90% of all chemical data analysed in this study were matched by predictions within data uncertainty. Results for cocaine and benzoylecgonine are similar, as expected for chemicals with expected similar pathways (cf. Fig. 3e–f and Fig. S25).

One explanation for the relatively high residual misfit for benzoylecgonine in subcatchments A and D is that data uncertainty is actually larger than that indicated by the spread of chemical concentrations measured at these localities. Another, more likely, explanation is that actual fluxes from neighbouring basins are not (cannot) be predicted accurately by the simple forward model, which does not consider temporally variable sources (e.g., input from intermittent CSOs). This issue is explored by first noting that there is a theoretical limit on observed downstream concentrations given the prescribed forward model.

Consider a two-component model, i.e., two measured concentrations d_1 (upstream) and d_2 (downstream) with associated unique subcatchments with concentrations c_1 and c_2, and fluxes q_1 and q_2. Downstream concentrations measured at the sample locations will be $d_1 = c_1$ and $d_2 = (c_1q_1 + c_2q_2)/(q_1 + q_2)$, thus

$$c_2 = d_2\left(\frac{q_1 + q_2}{q_2}\right) - d_1\left(\frac{q_1}{q_2}\right). \tag{16}$$

As concentrations and fluxes must be ≥ 0, $d_2(q_1 + q_2)/q_1 \geq d_1$, which places limits on the possible values of concentrations that can be predicted. For example, for equal fluxes, $d_2 \geq d_1/2$, i.e., concentrations at locality d_2 cannot be less than half the concentration at d_1. More generally:

$$\frac{d_{i+1}}{d_i} \geq \frac{\Sigma_{j=1}^{i} q_j}{\Sigma_{j=1}^{i+1} q_j}. \tag{17}$$

If these inequalities are not satisfied, it could indicate that the forward model is not appropriate, i.e., processes other than conservative mixing are important or that calculated fluxes are incorrect.

High calculated concentrations of most chemicals in subcatchment D almost always satisfy the inequality in all iterations of the Monte Carlo experiment. In other words, d_D/d_C is usually $> (q_A + q_B + q_C)/(q_A + q_B + q_C + q_D)$. These results emphasise the importance of regularising the inverse model, which utilises a simple forward model that does not include complex hydrology or chemical transformations.

The results summarised above demonstrate that the inverse model can effectively identify and quantify sources of chronic pollution. They also suggest that for chemicals with rapidly fluctuating concentrations, such as those related to CSOs, the primary source of benzoylecgonine in waterways as noted by Munro et al. (2019); Rapp-Wright et al. (2022), it may be necessary to incorporate time-varying effects to better explain the observed data. Enhancing the temporal resolution of sampling campaigns would be one way to better track the dynamic patterns of such contaminants.

Misfit between observed and predicted concentration at localities A and D appears to arise from two sources. First, added noise can generate small unrealistic relative changes in concentration downstream (such that the above inequality is not satisfied). Second, accurately constraining fluxes (q; i.e., effective drainage areas) is crucial to reducing uncertainty. Decreasing the run-off from subcatchment A reduces the residual misfit between observed and theoretical concentrations at sampling localities A–J (Fig. S36–S37 & S42).

Results from inverse modelling of all other chemicals are shown in Fig. S20–32 and summarised in Fig. 4. The source apportionment calculations indicated that subcatchment D was a dominant source of pesticides (imidacloprid and acetamiprid) in the R. Wandle. Similarly, this subcatchment was a dominant source of the pharmaceuticals sulfamethoxazole, trimethoprim, carbamazepine, diclofenac, tramadol, and venlafaxine. In contrast, calculated concentrations of the antibiotics azithromycin and clarithromycin, salicylic acid, cocaine and its

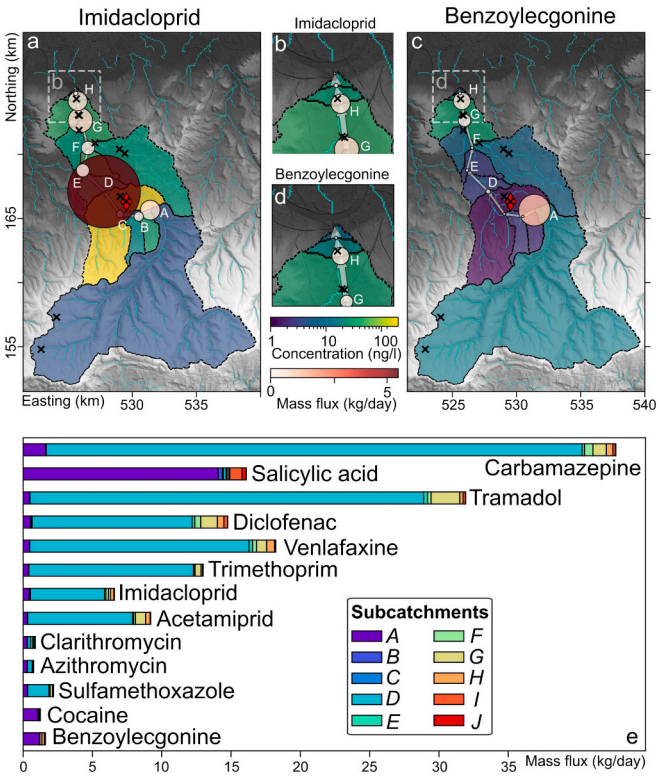


Fig. 5. Calculated fluxes of contaminants from Wandle subcatchments. (a–d) Coloured regions = subcatchments A–J (source areas; see Fig. 3) coloured by calculated concentrations of insecticide imidacloprid and cocaine's metabolite benzoylecgonine. Circles centred at sample localities = fluxes from corresponding subcatchments (circle area is proportional to flux). (a & d) Concentrations and fluxes close to Thames confluence. (e) Bar heights = cumulative mass fluxes (kg/day) of all chemicals inverted in this study; coloured sub-bars = contributions from each subcatchment (see key and Fig. 1). Note dominant sources of high chemical flux: subcatchment D (e.g., carbamazepine, tramadol, diclofenac, venlafaxine, trimethoprim, acetamiprid) and subcatchment A (e.g., salicylic acid). Calculated concentrations for all contaminants are shown in Fig. S20–S32 and summarised in Fig. 4.

metabolite benzoylecgonine are relatively low in subcatchment D. Instead, they tend to be higher close to the R. Wandle's confluence with the R. Thames. There are seven known CSO outlets in subcatchments downstream of D, which could be sources of untreated wastewater explaining the high concentrations in these subcatchments. In general, septic tanks, food processing plants, urban storm runoff and landfill sites can also be other important sources of contaminants. These results indicate that different contaminants in the R. Wandle catchment were derived from distinct sources upstream. However, they also emphasise that subcatchment D was the dominant source for a large number of contaminants. This catchment includes the Beddington wastewater treatment plant and Beddington Farmlands, a landfill site currently undergoing restoration. These results indicate that the locations of suspected point sources could be independently recovered just from downstream measurements at carefully chosen locations. We suggest that the capability and reach of routine monitoring at larger scales could be significantly enhanced by prioritising selected sites to support such modelling approaches.

3.3. Chemical fluxes

Fig. 5a–d shows the calculated fluxes of imidacloprid and benzoylecgonine from subcatchments A–J. Imidacloprid flux is dominated by subcatchment D. In contrast, the efflux of benzoylecgonine from subcatchment D is low, whereas subcatchments A and G–J have relatively

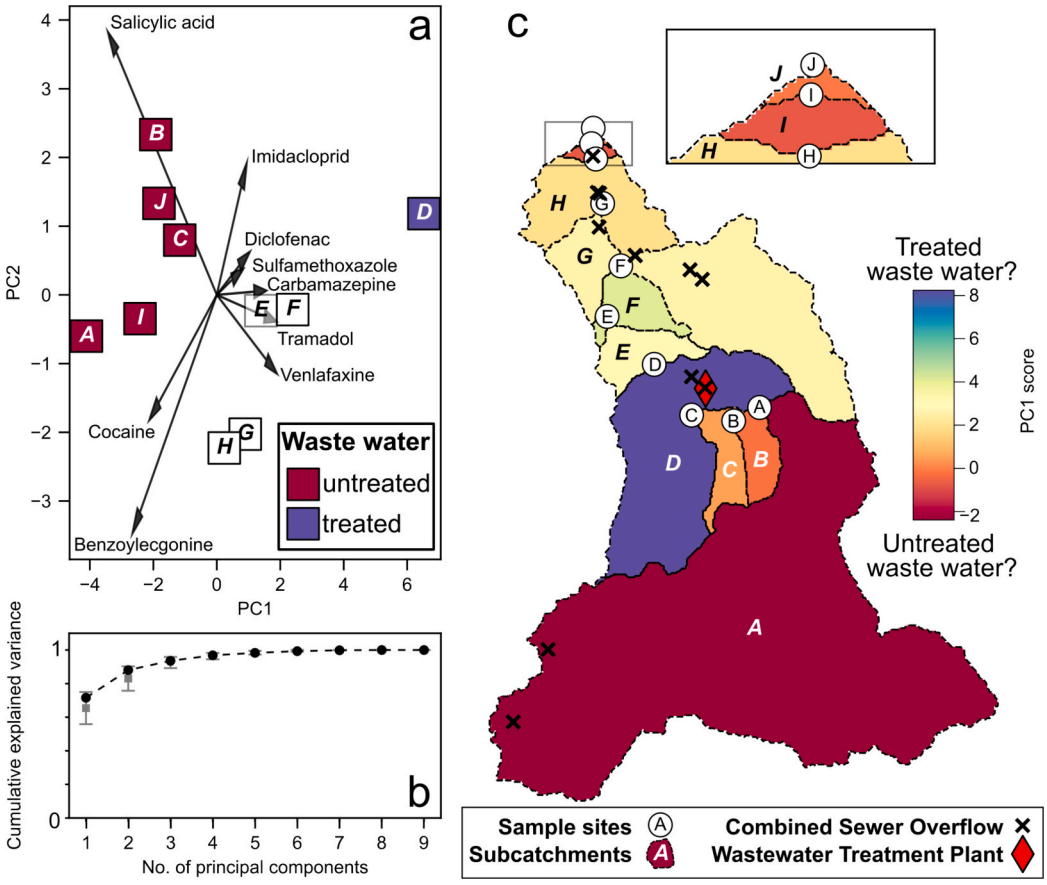


Fig. 6. Principal component analysis of source (subcatchment) contaminant concentrations in the Wandle. (a) Biplot showing chemical loading of the first two principal components explaining >85 % of variance of subcatchment contaminant concentrations. Colours indicate interpreted dominant pollutant pathway in each subcatchment. (b) Cumulative variance as a function of number of principal components. Black = results shown in panels (a) and (c); acetamiprid, trimethoprim, clarithromycin and azithromycin are not included in the analyses. Grey squares and error bars = mean and 2σ for all 715 combinations of removed chemicals (see body text for details). (c) Map showing scores for the first principal component in each subcatchment (see Fig. 1). Note distribution of interpreted treated and untreated waste water. Inset shows zoom at confluence with Thames river. White labeled circles = sample localities; x = CSOs, red diamond = Beddington WWTP.

high fluxes. Fig. 5e shows calculated fluxes for all chemicals and subcatchments examined. It demonstrates that subcatchment D dominates the flux of carbamazepine, tramadol, diclofenac, venlafaxine, trimethoprim, imidacloprid, acetamiprid and sulfamethoxazole in the Wandle. In contrast, subcatchment A is an important source of salicylic acid, cocaine and benzoylecgonine. In fact, calculated fluxes indicate that subcatchments A and E–J are important sources of most chemicals, although for subcatchment A this is in part related simply to its larger size relative to other subcatchments. Associated risk quotients and uncertainties are summarised in Fig. S33.

3.4. Principal component analysis

Fig. 6 shows the results of the PCA of calculated contaminant concentrations, used to establish relationships between sources of chemicals in the R. Wandle catchment. We tested all 715 combinations of excluded chemicals in a bootstrap analysis. The PCA results shown in Fig. 6 exclude acetamiprid, trimethoprim, clarithromycin, and azithromycin simply because they have relatively low and similar concentrations. Panel (a) shows the first two principal components of the nine remaining chemicals and the subcatchments. More than 70% of the variance was explained by the first component (PC1), > 85% was explained by the first two components. As expected, cocaine and its metabolite benzoylecgonine had similar loadings on these two components (see vectors in panel a). Salicylic acid had a similarly negative loading on the first principal component. In contrast, the pharmaceuticals (including pain medications, antibiotics and pesticides) had positive loadings on the first principal component. Subcatchment D is significantly positively loaded onto the first principal component. Other subcatchments are negatively loaded, or close to zero. As the first principal component contains the majority of the variability in the dataset, it is visualised spatially in panel (c). This map shows the loadings on the first principal component for each subcatchment. These results emphasise that the contaminant composition of subcatchment D was distinct from the other subcatchments.

These results are largely unaffected by the choice of removed chemicals. For instance, mean scores for the first principal components from the bootstrap analyses of the 715 combinations are consistent with Fig. 6c and associated standard deviations are relatively low (see Fig. 6b).

These results were interpreted as indicating the presence of two primary pathways for chemical contaminants into the R. Wandle. The contaminant load from subcatchment D is fundamentally distinct from the contaminant load from any other subcatchment. Positive loadings on the first principal component were interpreted as representing the 'fingerprint' of treated wastewater. These results could be explained by the presence of the Beddington WWTP in subcatchment D. In contrast, subcatchments with negative loadings on PC1, i.e., A, B, C, I, J, were interpreted as having a distinct contaminant pathway. This source was interpreted as corresponding to *untreated* wastewater due to the presence of illicit drugs, which are present in wastewater and efficiently removed if they reach wastewater facilities (Munro et al., 2019). These

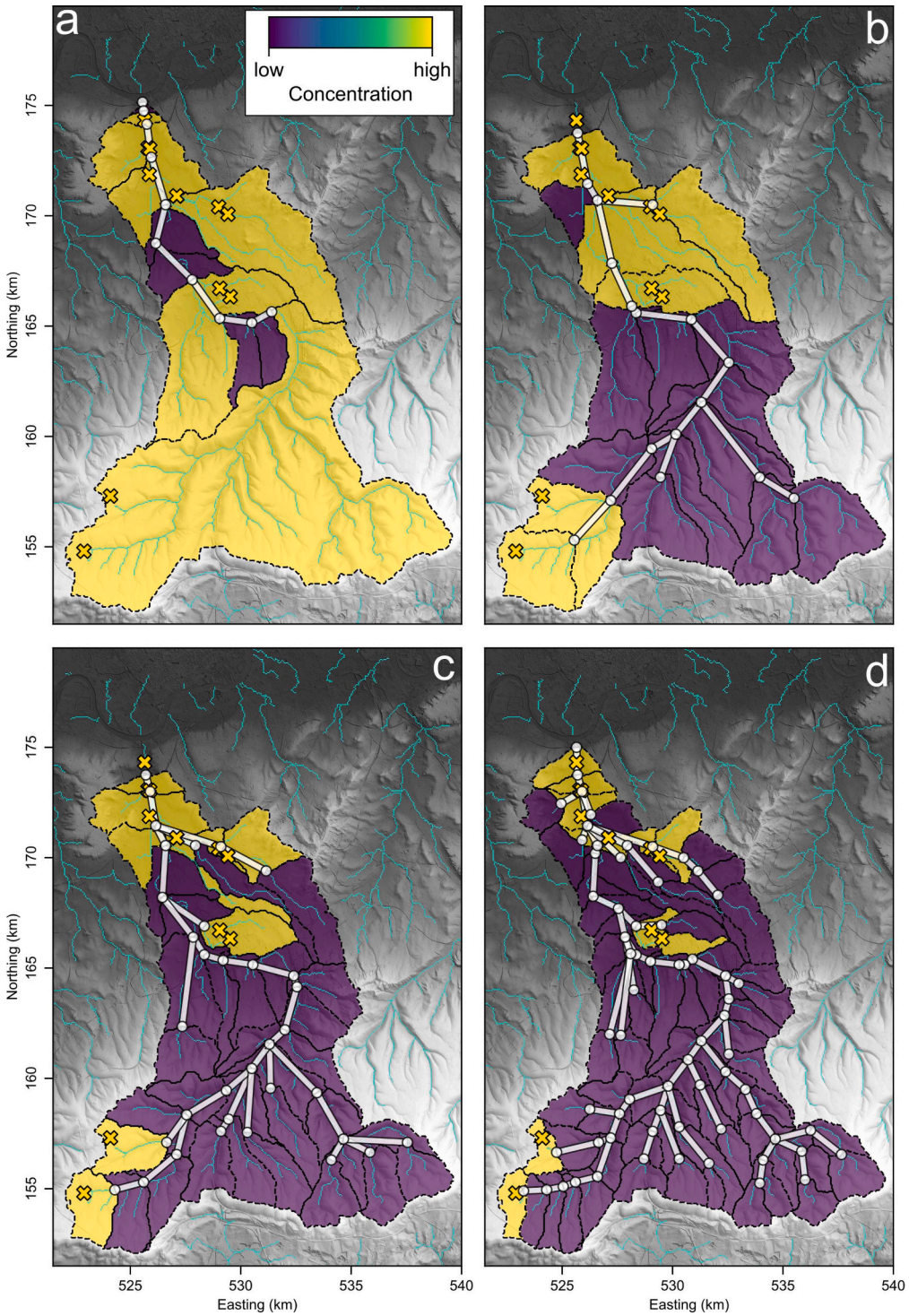


Fig. 7. Improving precision of source apportionment. (a) Apportionment using actual sampling localities from this study (see Fig. 1). (b–d) Apportionment using sampling localities distributed so that unique subcatchments have equal area. (b) 17 samples; 5 km^2. (c) 34 samples, 2 km^2. (d) 66 samples, 1 km^2. Yellow crosses = locations of synthetic point sources of contaminants (e.g., note that they are located at actual CSO events). White circles = sample sites; white lines = network connecting sample sites. Turquoise curves = drainage network. Coloured regions with dashed outlines = subcatchments unique to each sample, coloured by calculated source concentration.

results are broadly consistent with hierarchical cluster analysis of contaminants in rivers across London, which emphasises commonalities between contaminant concentrations measured downstream of WWTPs, and commonalities of concentrations measured in catchments without upstream WWTPs (Egli et al., 2023). We emphasise that our analysis is distinct from this previous analysis as we are directly analysing the calculated *sources* rather than observed concentrations *downstream*. Untreated wastewater could enter the river via sewerage misconnections, CSOs, landfill leaching, runoff including urban storm-water or perhaps via direct dumping (e.g., Masoner et al., 2016, 2019; Hubbard et al., 2022; Webb et al., 2021). CSOs may have been active in some of the subcatchments during the period of sampling (see Fig. S4–S6), although specific discharge data were not available. Subcatchments E, F, G, H were not found to be compositionally distinct sources of any

contaminants.

3.5. Optimal sampling strategies

Finally, optimisation of source recovery in future sampling campaigns was examined (see Sect. 2.10). Fig. 7 demonstrates how sampling strategies that prioritise equal-area apportionment could yield assessments of chemical concentrations in sources areas unbiased by preconceived pollutant sources. This figure shows the results from a suite of inverse models solved using different distributions of sampling localities (white circles) for an arbitrary (well mixed, conservative) contaminant. The yellow crosses indicate point sources of the contaminant throughout the Wandle catchment (e.g., CSO), which were constant in each test. Chemical concentrations at sampling localities were calculated using the forward model (Eq. 9). Panel (a) shows the results from the first test in which 'measured' concentrations at the actual sampling localities used in this study were inverted (see e.g., Fig. 1). Panels (b–d) show the results from inverting an increasingly high number of samples located so that subcatchments upstream had an equal area. These results indicate that sources of contaminants could be more precisely, and objectively, identified using a sampling strategy optimised to determine contaminant concentrations in equal-area subcatchments. In other words, sampling campaigns that seek to determine sources of chemicals, especially those that are unknown or suspect, should incorporate knowledge of network topology when deciding where to collect samples for chemical analyses.

4. Discussion

4.1. An objective approach to determining contaminant provenance

A principal finding from this study is that spot measurements of contaminant concentrations in water can be combined with the topology of waterway networks to invert for sources of chemicals. We emphasise that this approach is objective and that it solves for actual sources of contaminants (rather than, say, the magnitude of a priori assumed sources) without recourse to additional information (e.g., satellite imagery, land use, locations of point-sources). The general framework developed in this study could prove useful for identifying locations to build WWTPs and in environmental risk assessment for infrastructure planning. Ecotoxicologically relevant chemicals in waterways, including insecticides, pharmaceuticals and illicit drugs, could be apportioned to sources upstream with no prior knowledge of provenance. The results indicate that conservative mixing, assumed in this study, can provide enough complexity to accurately predict concentrations of most contaminants downstream (e.g., Fig. 3). In other words, complex hydrological modelling or chemical transformation may not always be required to reliably predict concentrations of contaminants in waterways downstream. However, the inverse scheme developed has sufficient flexibility such that more complex forward models (e.g., allowing for chemical decay) could be straightforwardly incorporated if necessary.

4.2. Contaminant sources in the River Wandle

Calculated source apportionment and chemical fluxes for the Wandle emphasised the importance of subcatchment D as a source of the pesticides and of most pharmaceuticals investigated. Efflux from the Beddington WWTP appeared to be the most likely source of these chemicals. In other words, it would appear that these chemicals, directed to this sewer network terminus from the Beddington WWTP catchment (incorporating Croydon, Carshalton, Coulsdon, Caterham, and Warlingham), evaded treatment and exited the plant at concentrations likely to have adverse ecological impacts (Fig. 3 & S33). These results are consistent with independent chemical analysis of treated and untreated effluent from other areas (Du et al., 2014; Munro et al., 2019; Golovko et al., 2021; Rapp-Wright et al., 2022). An obvious implication of this finding is that reducing contaminant concentrations in subcatchment D could be expected to significantly reduce exposure to many chemicals downstream. We note that strategies to reduce contaminant concentrations in single subcatchments (or single watersheds) may result in transformation products with their own ecotoxicological impacts and that disposal can have adverse impacts elsewhere (Hu et al., 2023; Löffler et al., 2023).

In contrast with the apportionment of these pesticides and most of the pharmaceuticals studied in the Wandle catchment, sources of illicit drugs and salicylic acid were not dominated by a single subcatchment. They appeared to have had sources in multiple subcatchments. Consider that the first principal component of calculated source concentrations divides the Wandle catchment into two groups of subcatchments of concern, one is dominated by subcatchment D (and by inference the Beddington WWTP), and the other by more dispersed entry (e.g., CSO spills, sewer leaks, runoff, direct dumping). Reported rainfall and river flow rates were low during the acquisition of chemical measurements at localities A–J. A tentative comparison with more recent, and more complete, data suggests that CSO discharges may also have been low then (see Fig. S5–S6). However, information about dry weather discharges, and time-series of CSO spills during the 2019–2021 sampling period, is not available. A denser or more strategically sited sampling campaign could help determine the source(s) of these chemicals. We note also that the nested ANOVA suggested that temporally varying sources may be important for these chemicals, indicating that time-integrated sampling may be beneficial. Options for improving source apportionment are explored in Sect. 4.3.

4.3. Optimal sampling strategies

The chemical concentrations inverted in this study were from a legacy dataset that was not collected with inverse source apportionment in mind. Future sampling strategies could be optimised for the source apportionment methods presented here. For example, our analysis identified subcatchment A as a potential source of illicit drugs in the R. Wandle. Unfortunately, this particular subcatchment covers a large area. As such, ascribing sources to a specific pollutant pathway, for example, a specific CSO or a misconnection, is severely limited by the distribution of sampling downstream.

Optimal sampling combined with the inverse approach developed in this study could provide practical means to identify and quantify contaminant sources throughout drainage networks (Fig. 7). We note that the samples inverted in this study were collected in autumn and winter. A fruitful avenue for future work could combine optimal sampling strategies with sampling at different times to assess spatio-temporal changes in sources of contaminants in response to environmental (e.g., climate) and socio-economic change. Finally, assessing how concentrations of chemicals respond to tidal influences (e.g., along the tidal reach of the Thames) is likely to be important for apportioning chemicals to sources along large rivers.

5. Conclusion

Contaminant concentrations measured in waterways provide crucial information about pollutant sources upstream. Such information is central to mitigating contaminant risks for ecology and human health. In this study, an inverse modelling strategy was applied for the first time to an urban river, the R. Wandle, a tributary of the R. Thames, UK, to recover sources of thirteen chemical contaminants (including pesticides, pharmaceuticals and illicit drugs) measured along its drainage network.

The recoverability of concentrations in the Wandle's source areas was successfully demonstrated by first inverting synthetic data. The concentrations of the thirteen contaminants, measured at ten localities along the R. Wandle in 2020–2021, were then inverted. The inverse approach demonstrated how data and model uncertainties could be

incorporated into source apportionment. Two distinct chemical contaminant pathways into the river were identified. The first was the Beddington wastewater treatment plant, which appeared to be the dominant source of the pharmaceuticals and pesticides measured in this study. The second pathway was more dispersed, indicating multiple points of entry, particularly for cocaine and its metabolite benzoylecgonine and some pharmaceuticals (e.g., salicylic acid). We tentatively suggest that this second pathway corresponds to untreated wastewater from sewerage misconnections, combined sewer overflow events, runoff or perhaps direct dumping.

By combining recovered source concentrations with river discharge data, the mass fluxes of all studied chemicals were calculated. To support modelling efforts, including improving the precision of source apportionment, more efficient sampling strategies were shown to be required. An objective method to identify potential future sample sites unbiased by preconceptions of pollutant sources was demonstrated. Combining optimal sampling strategies and inverse modelling of measured contaminant concentrations could identify the location and concentrations of contaminant sources in a wide variety of waterways at significantly larger scales.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2024.172827.

CRediT authorship contribution statement

Kajetan Chrapkiewicz: Writing – review & editing, Formal analysis, Methodology, Software, Visualization, Writing – original draft. **Alex G. Lipp:** Writing – review & editing, Visualization, Validation, Software, Methodology, Conceptualization. **Leon P. Barron:** Writing – review & editing, Funding acquisition, Data curation, Conceptualization. **Richard Barnes:** Writing – review & editing, Software, Methodology. **Gareth G. Roberts:** Writing – review & editing, Writing – original draft, Visualization, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Data availability

Code to perform the source apportionment calculations (the inverse model) is available from https://github.com/kmch/SourceApp. Code to identify localities that provide equal area sampling of a drainage network is available at zenodo.org/records/7311352. LIDAR digital elevation data used is available from https://environment.data.gov.uk/dataset/ce8fe7e7-bed0-4889-8825-19b042e128d2. HydroSHEDS drainage data is available from https://www.hydrosheds.org. CSO event data is available from https://www.thameswater.co.uk/about-us/performance/river-health/storm-discharge-data#annual-edm-reports. Environment Agency data for the Wandle catchment can be accessed at https://environment.data.gov.uk/catchment-planning/OperationalCatchment/3514. Weather data was accessed at http://nw3weather.co.uk/. The NORMAN ecotoxicology database that contains *predicted no-effect concentrations* of chemicals can be accessed at https://www.norman-network.com/nds/ecotox. PCA was performed using the Python scikit library sklearn.decomposition. PCA (Pedregosa et al., 2011).

Acknowledgements

We thank the Natural Environment Research Council for funding this work (NE/X010805/1). AGL is funded by a Junior Research Fellowship from Merton College, Oxford. LB is part-funded under the National Institute for Health and Care Research (NIHR) Health Protection Research Units in Chemical and Radiation Threats and Hazards, as well as Environmental Exposures and Health, both partnerships between the UK Health Security Agency and Imperial College London. The views expressed are those of the authors and not necessarily those of the NIHR, UK Health Security Agency or the Department of Health and Social Care.

References

Aitchinson, J., 1983. Principal component analysis of compositional data. Biometrika 70 (1), 57–65.
Aster, R.C., Borchers, B., Thurber, C.H., 2005. Parameter estimation and inverse problems, 1st edition. Elsevier.
Barnes, R., Lipp, A.G., 2024. Identifying tracer and pollutant sources in drainage networks from point concentration observations using an efficient convex unmixing scheme. Water Resour. Res. https://doi.org/10.1029/2023WR036159.
Bradley, P.M., Journey, C.A., Romanok, K.M., Barber, L.B., Buxton, H.T., Foreman, W.T., Furlong, E.T., Glassmeyer, S.T., Hladik, M.L., Iwanowicz, L.R., Jones, D.K., Kolpin, D.W., Kuivila, K.M., Loftin, K.A., Mills, M.A., Meyer, M.T., Orlando, J.L., Reilly, T.J., Smalling, K.L., Villeneuve, D.L., 2017. Expanded target-chemical analysis reveals extensive mixed-organic-contaminant exposure in U.S. streams. Environ. Sci. Technol. 51 (9), 4792–4802.
Campos-Mañas, M.C., Plaza-Bolaños, P., Sánchez-Pérez, J.A., Malato, S., Agüera, A., 2017. Fast determination of pesticides and other contaminants of emerging concern in treated wastewater using direct injection coupled to highly sensitive ultra-high performance liquid chromatography-tandem mass spectrometry. J. Chromatogr. A 1507, 84–94.
Carraro, L., Mächler, E., Wüthrich, R., Altermatt, F., 2020. Environmental DNA allows upscaling spatial patterns of biodiversity in freshwater ecosystems. Nat. Commun. 11 (1), 3585.
Carraro, L., Stauffer, J.B., Altermatt, F., 2021. How to design optimal edna sampling strategies for biomonitoring in river networks. Environ. DNA 3 (1), 157–172.
Chen, K., Meng Liu, Q., Hua Peng, W., Liu, Y., Tao Wang, Z., 2023. Source apportionment of river water pollution in a typical agricultural city of Anhui Province, eastern China using multivariate statistical techniques with APCS–MLR. Water Science and Engineering 16 (2), 165–174.
Christophersen, N., Neal, C., Hooper, R.P., Vogt, R.D., Andersen, S., 1990. Modelling streamwater chemistry as a mixture of soilwater end-members — a step towards second-generation acidification models. J. Hydrol. 116 (1), 307–320.
Collins, A.L., Blackwell, M., Boeckx, P., Chivers, C.-A., Emelko, M., Evrard, O., Foster, I., Gellis, A., Gholami, H., Granger, S., Harris, P., Horowitz, A.J., Laceby, J.P., Martinez-Carreras, N., Minella, J., Mol, L., Nosrati, K., Pulley, S., Silins, U., da Silva, Y.J., Stone, M., Tiecher, T., Upadhayay, H.R., Zhang, Y., 2020. Sediment source fingerprinting: benchmarking recent outputs, remaining challenges and emerging themes. J. Soils Sediments 20, 4160–4193.
Comber, S.D.W., Smith, R., Daldorph, P., Gardner, M.J., Constantino, C., Ellor, B., 2013. Development of a chemical source apportionment decision support framework for catchment management. Environ. Sci. Technol. 47 (17), 9824–9832.
Couchman, L., Fisher, D.S., Subramaniam, K., Handley, S.A., Boughtflower, R.J., Benton, C.M., Flanagan, R.J., 2018. Ultra-fast lc–ms/ms in therapeutic drug monitoring: quantification of clozapine and norclozapine in human plasma. Drug Test Anal. 10 (2), 323–329.
Cox, B.A., 2003. A review of currently available in-stream water-quality models and their applicability for simulating dissolved oxygen in lowland rivers. Sci. Total Environ. 314–316.
Dixon, W., Smyth, G.K., Chiswell, B., 1999. Optimized selection of river sampling sites. Water Res. 33 (4), 971–978.
Domahidi, A., Chu, E., Boyd, S., 2013. ECOS: an SOCP solver for embedded systems. In: 2013 European Control Conference (ECC).
Du, B., Price, A.E., Scott, W.C., Kristofco, L.A., Ramirez, A.J., Chambliss, C.K., Yelderman, J.C., Brooks, B.W., 2014. Comparison of contaminants of emerging concern removal, discharge, and water quality hazards among centralized and on-site wastewater treatment system effluents receiving common wastewater influent. Sci. Total Environ. 466-467, 976–984.
Egli, M., Hartmann, A., Rapp Wright, H., Ng, K.T., Piel, F.B., Barron, L.P., 2021. Quantitative determination and environmental risk assessment of 102 chemicals of emerging concern in wastewater-impacted rivers using rapid direct-injection liquid chromatography—tandem mass spectrometry. Molecules 26 (18).
Egli, M., Rapp-Wright, H., Oloyede, O., Francis, W., Preston-Allen, R., Friedman, S., Woodward, G., Piel, F.B., Barron, L.P., 2023. A One-Health environmental risk assessment of contaminants of emerging concern in London's waterways throughout the SARS-CoV-2 pandemic. Environ. Int. 180 (108210), 108210.
Feraud, M., O'Brien, J.W., Samanipour, S., Dewapriya, P., van Herwerden, D., Kaserzon, S., Wood, I., Rauert, C., Thomas, K.V., 2023. Inspectra – a platform for identifying emerging chemical threats. J. Hazard. Mater. 455, 131486.
Golovko, O., Örn, S., Sörengård, M., Frieberg, K., Nassazzi, W., Lai, F.Y., Ahrens, L., 2021. Occurrence and removal of chemicals of emerging concern in wastewater treatment plants and their impact on receiving water systems. Sci. Total Environ. 754, 142122.
Hobley, D.E.J., Adams, J.M., Nudurupati, S.S., Hutton, E.W.H., Gasparini, N.M., Istanbulluoglu, E., Tucker, G.E., 2017. Creative computing with Landlab: an open-source toolkit for building, coupling, and exploring two-dimensional numerical models of Earth-surface dynamics. Earth Surf. Dyn. 5 (1), 21–46.

Hu, J., Lyu, Y., Chen, H., Li, S., Sun, W., 2023. Suspect and nontarget screening reveal the underestimated risks of antibiotic transformation products in wastewater treatment plant effluents. Environ. Sci. Technol. 57 (45), 17439–17451.
Hubbard, L.E., Kolpin, D.W., Givens, C.E., Blackwell, B.R., Bradley, P.M., Gray, J.L., Lane, R.F., Masoner, J.R., McCleskey, R.B., Romanok, K.M., Sandstrom, M.W., Smalling, K.L., Villeneuve, D.L., 2022. Food, beverage, and feedstock processing facility wastewater: a unique and underappreciated source of contaminants to u.s. streams. Environ. Sci. Technol. 56 (2), 1028–1040.
Linke, S., Lehner, B., Ouellet Dallaire, C., Ariwi, J., Grill, G., Anand, M., Beames, P., Burchard-Levine, V., Maxwell, S., Moidu, H., Tan, F., Thieme, M., 2019. Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. Scientific Data 6.
Lipp, A.G., Roberts, G.G., Whittaker, A.C., Gowing, C.J.B., Fernandes, V.M., 2020. River sediment geochemistry as a conservative mixture of source regions: observations and predictions from the Cairngorms, UK. J. Geophys. Res. Earth 125 (12).
Lipp, A.G., Roberts, G.G., Whittaker, A.C., et al., 2021. Source region geochemistry from unmixing downstream sedimentary elemental compositions. Geochem. Geophys. Geosyst. 22, e2021GC009838 https://doi.org/10.1029/2021GC009838.
Löffler, P., Escher, B.I., Baduel, C., Virta, M.P., Lai, F.Y., 2023. Antimicrobial transformation products in the aquatic environment: global occurrence, ecotoxicological risks, and potential of antibiotic resistance. Environ. Sci. Technol. 57 (26), 9474–9494.
Masoner, J.R., Kolpin, D.W., Furlong, E.T., Cozzarelli, I.M., Gray, J.L., 2016. Landfill leachate as a mirror of today's disposable society: pharmaceuticals and other contaminants of emerging concern in final leachate from landfills in the conterminous United States. Environ. Toxicol. Chem. 35 (4), 906–918.
Masoner, J.R., Kolpin, D.W., Cozzarelli, I.M., Barber, L.B., Burden, D.S., Foreman, W.T., Forshay, K.J., Furlong, E.T., Groves, J.F., Hladik, M.L., Hopton, M.E., Jaeschke, J.B., Keefe, S.H., Krabbenhoft, D.P., Lowrance, R., Romanok, K.M., Rus, D.L., Selbig, W. R., Williams, B.H., Bradley, P.M., 2019. Urban stormwater: an overlooked pathway of extensive mixed contaminants to surface and groundwaters in the United States. Environ. Sci. Technol. 53 (17), 10070–10081.
Milledge, D.G., Lane, S.N., Heathwaite, A.L., Reaney, S.M., 2012. A Monte Carlo approach to the inverse problem of diffuse pollution risk in agricultural catchments. Sci. Total Environ. 433, 434–449.
Morris, D.G., Flavin, R.W., 1994. Sub-set of the UK 50 m by 50 m Hydrological Digital Terrain Model Grids. NERC, Institute of Hydrology, Wallingford.
Munro, K., Martins, C.P.B., Loewenthal, M., Comber, S., Cowan, D.A., Pereira, L., Barron, L.P., 2019. Evaluation of combined sewer overflow impacts on short-term pharmaceutical and illicit drug occurrence in a heavily urbanised tidal river catchment (London, UK). Sci. Total Environ. 657, 1099–1111.
O'Callaghan, J.F., Mark, D.M., 1984. The extraction of drainage networks from digital elevation data. Computer Vision, Graphics, and Image Processing 28 (3), 323–344.
O'Donoghue, B., Chu, E., Parikh, N., Boyd, S., 2016. Conic optimization via operator splitting and homogeneous self-dual embedding. J. Optim. Theory Appl. 169.
Palma, P., Köck-Schulmeyer, M., Alvarenga, P., Ledo, L., Barbosa, I., López de Alda, M., Barceló, D., 2014. Risk assessment of pesticides detected in surface water of the alqueva reservoir (guadiana basin, southern of Portugal). Sci. Total Environ. 488–489, 208–219.
Parker, R.L., 1994. Geophysical Inverse Theory, vol 1. Princeton University Press.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Perkins, R., Barron, L., Glauser, G., Whitehead, M., Woodward, G., Goulson, D., 2024. Down-the-drain pathways for fipronil and imidacloprid applied as spot-on parasiticides to dogs: estimating aquatic pollution. Sci. Total Environ. 917, 170175.
Rapp-Wright, H., Regan, F., White, B., Barron, L.P., 2022. A year-long study of the occurrence and risk of over 140 contaminants of emerging concern in wastewater influent, effluent and receiving waters in the republic of Ireland. Sci. Total Environ. 860 (160379), 1–12.
Reemtsma, T., Alder, L., Banasiak, U., 2013. A multimethod for the determination of 150 pesticide metabolites in surface water and groundwater using direct injection liquid chromatography–mass spectrometry. J. Chromatogr. A 1271 (1), 95–104.
Richardson, A.K., Chadha, M., Rapp-Wright, H., Mills, G.A., Fones, G.R., Gravell, A., Stürzenbaum, S., Cowan, D.A., Neep, D.J., Barron, L.P., 2021. Rapid direct analysis of river water and machine learning assisted suspect screening of emerging contaminants in passive sampler extracts. Anal. Methods 13, 595–606.
Shapiro, N., Ritzwoller, M., 2002. Monte-Carlo inversion for a global shear-velocity model of the crust and upper mantle. Geophys. J. Int. 151 (1), 88–105.
Singhal, A., Jaseem, M., Divya, Sarker, S., Prajapati, P., Singh, A., Jha, S.K., 2023. Identifying potential locations of hydrologic monitoring stations based on topographical and hydrological information. Water Resour. Manag. 38, 369–384.
Taylor, J.R., 1997. An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements, 2 edition. University Science Books.
Visser, A., Beevers, L., Patidar, S., 2019. The impact of climate change on hydroecological response in chalk streams. Water 11 (3).
Webb, D.T., Zhi, H., Kolpin, D.W., Klaper, R.D., Iwanowicz, L.R., LeFevre, G.H., 2021. Emerging investigator series: municipal wastewater as a year-round point source of neonicotinoid insecticides that persist in an effluent-dominated stream. Environ Sci Process Impacts 23, 678–688.
Weltje, G.J., 1997. End-member modeling of compositional data: numerical-statistical algorithms for solving the explicit mixing problem. Math. Geol. 29, 503–549.
Zhi, H., Kolpin, D.W., Klaper, R.D., Iwanowicz, L.R., Meppelink, S.M., LeFevre, G.H., 2020. Occurrence and spatiotemporal dynamics of pharmaceuticals in a temperate-region wastewater effluent-dominated stream: variable inputs and differential attenuation yield evolving complex exposure mixtures. Environ. Sci. Technol. 54 (20), 12967–12978.
Zhou, G., Chen, S., Li, A., Xu, C., Jing, G., Chen, Q., Hu, Y., Tang, S., Lv, M., Xiao, K., 2023. Pollution source apportionment of river tributary based on pmf receptor model and water quality remote sensing in xinjian river, China. Water 15 (1).