





## A validation workflow for treatment wetland performance data

Sophie Hai Yen Guillaume-Ruty <sup>a</sup>, Josep Pueyo-Ros <sup>b</sup>, Joaquim Comas <sup>b,c</sup>  
and Nicolas Forquet <sup>a,\*</sup>

<sup>a</sup> Research Unit REVERSAAL, INRAE, 5 rue de la Doua, Villeurbanne, France

<sup>b</sup> ICRA-CERCA, Emili Grahit, 101, 17003 Girona, Spain

<sup>c</sup> LEQUIA, University of Girona, C/M<sup>a</sup> Aurèlia Capmany, 69, 17003 Girona, Spain

\*Corresponding author. E-mail: nicolas.forquet@inrae.fr

 SHYG, 0009-0002-4038-7409; JP, 0000-0002-1236-5651; JC, 0000-0002-5692-0282; NF, 0000-0003-1154-5498

### ABSTRACT

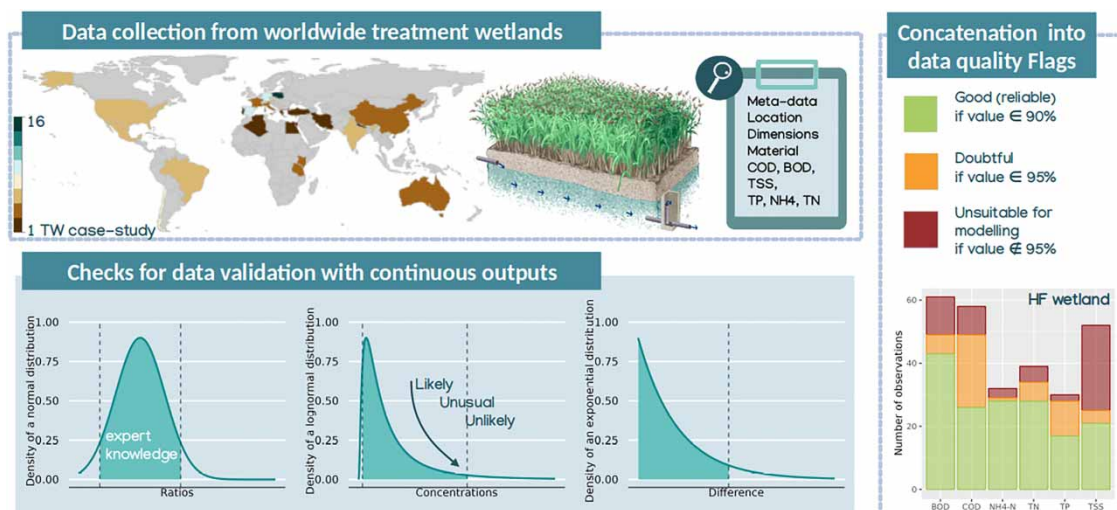
Treatment wetlands (TWs) effectively remove target pollutants and enhance urban water circularity and resilience. They constitute a prominent solution for urban wastewater treatment, thanks to their adaptability across various types of wastewater, scales and climatic conditions. However, the disparity in TW designs and the focus on a restricted set of variables applicable to research studies impede any comprehensive evaluation and comparison of TW performance. Our study introduces a methodology for data validation, in concurrently establishing a workflow specific to TW. This approach is aimed at defining the scope and relationships within the data, implementing checks and concatenating them into a quality flag, as an initial step towards building reliable statistical models. We underscore the importance of both mobilising comprehensive knowledge and identifying customary, yet implicit, choices intertwined in data processing. As for the application workflow, we collected and analysed data sourced from peer-reviewed papers on horizontal and vertical flow TW. Deficiencies were noted in key data elements like dimensions, concentrations and operational conditions. For the data analysis, relationships are highlighted between variables introduced for modelling purposes. These methodologies and workflows assess the quality of the data, in paving the way towards more dependable statistical models for TW design and implementation.

**Key words:** categorisation, data quality, methodology, pollutants, probabilistic, wastewater treatment

### HIGHLIGHTS

- We provide a six-step data validation methodology highlighting the key elements to consider, regardless of the field.
- We develop data quality checks, in proposing a probabilistic output that can be used in data-driven modelling.
- We analyse a worldwide-scale dataset on treatment wetlands that can serve as a ground reference for data-driven modelling.

## GRAPHICAL ABSTRACT



## INTRODUCTION

The field of water management is undergoing a shift in terms of its key challenges, transitioning from a fragmented, linear and technical perspective to a circular and interdisciplinary approach, i.e. in favour of Integrated Water Resources Management (IWRM) promoted by UN-EP (United Nations Environment Programme 2023). With goals such as cost limitation, resource conservation and enhancement of urban resilience, one promising solution lies in the adoption of nature-based solutions (NBS; Oral *et al.* 2020; Langergraber *et al.* 2021a, 2021b). These kinds of solutions encompass a range of actions and technologies leveraging and intensifying natural processes; more notably, they promote the following (Chan *et al.* 2018; Langergraber *et al.* 2021a, 2021b):

- Decentralised management practices, aimed at reducing energy consumption and relieving pressure on the sewer system.
- Onsite water reuse, for alleviating the strain on the demand for potable water.
- Urban greening initiatives, for mitigating heat islands, bolstering urban permeability to counter floods and enhancing the aesthetic appeal of urban landscapes.

Due to their versatile range of implementation scales, treatment wetlands (TWs) serve as a complement to conventional wastewater treatment plants. Moreover, they yield many additional advantages to pollutant removal, as highlighted in Masi *et al.* (2018). The limitation in TW adoption lies in the lack of standard design guidelines: rule-of-thumb and mass loading charts are two widely used methods, yet both remain over-simplified.

To facilitate TW implementation, a decision-support tool is under development within the MULTISOURCE European Project (<https://multisource.eu/>, based on Acuña *et al.* (2023)). In accordance with the paradigm shift towards an integrated approach, this tool aims to provide design suggestions and recommendations on the most suitable NBS implementation, in considering diverse criteria such as pollutant removal, costs and co-benefits. The approach for the design is the use of data-driven models. They are more flexible than previous methods given that they can capture (un)known and non-quantified mechanisms (refer to Supplementary Table S6). The design model is intended to establish the relationship between TW designs and treatment performance, by considering influential factors such as type of climate, wastewater composition and type of filtering substrate, as set forth in (Rousseau *et al.* 2004; Langergraber 2011). With a limited understanding of the influence of certain factors on pollutant removal processes, the development of data-driven models serves as a valuable *post-hoc* approach.

## Ensuring data reliability for modelling purposes

The reliability of primarily data-driven models hinges upon data quantity and quality, which necessitates focusing on the production of input data. When a model is trained on few case studies ( $n$  observations), in particular relative to the number of features  $p$ , an 'outlier' observation exerts a substantial influence on the model fit, and features wind up being inefficiently utilised. This scenario corresponds to a low degree of freedom (for a linear regression model error,  $v = n - p - 1$ ) or a low sample size-to-feature size ratio (SFR =  $n/p$ ) (Zhu *et al.* 2023). The risk of overfitting and low confidence in model parameters

from the lack of data can be mitigated by not using models requiring many (hyper)parameters, such as deep learning models and complex statistical models.

From a technical standpoint, the MULTISOURCE tool is designed to provide pre-dimensions of NBS specifically tailored to the context. Its scope is global, thus requiring the incorporation of a pre-sizing model based on a synthesis of TW case studies distributed across various geographic locations worldwide. For a global TW model, considerable variability in observations is expected, thus warranting consideration of an extensive set of features derived from a comprehensive range of possible variables. However, despite numerous scientific studies, few such models encompass broad scopes and furnish exhaustive datasets (Supplementary Table S4). Instead, they adopt localised scopes focusing on one or several TW systems and specific pollutants of interest. Their localised diversity complicates both data comparison and aggregation, thereby impeding effective pre-sizing and future implementation of TW systems. The construction of a dataset thus involves collecting data from heterogeneous sources, TW systems, operational routines and data formats. Unfortunately, much of this data remains difficult to mobilise and consequently stagnates as a 'data graveyard' (Corominas *et al.* 2018) due to flawed storage practices or incomplete technical reports (e.g. missing data, insufficient contextual or meta-data).

Given the challenges posed by a small and heterogeneous dataset, data validation constitutes a critical step in ensuring model reliability. Two methodologies have been developed specifically for time-series data, within the respective frameworks of activated sludge models (Rieger *et al.* 2010; Rieger 2012) and metrology in urban water management (Clemens-Meyer *et al.* 2021). These methodologies seek to execute several checks and categorise data quality. Their assessment of experimental data incorporates statistical tools and expert knowledge, while relying on five criteria, namely:

- **Plausibility:** Measurements need to be in line with anticipated outcomes under specified conditions as well as adhere to the principles of physics and measurement instrument ranges.
- **Consistency:** Repeated and time-series data need to be consistent with one another; moreover, the measurements recorded by different devices must maintain concordance without displaying any contradiction.
- **Accuracy:** The measurements need to lie close to the expected 'true' value. Accuracy is also sometimes assessed in conjunction with precision, through assessing the closeness of repeated measurements.
- **Auditability:** The process of data acquisition and reconciliation needs to be transparent and traceable. Comprehensive documentation including date, location, experimental design and instrument details enables reproducibility of the study.
- **Synchronicity:** The measurements obtained from different devices need to be coordinated in time.

### Decision making to identify outliers

Data production involves numerous decision points, ranging from experimental set-up to measurement techniques and reconciliation practices. Data validation entails the detection of faulty values and outliers. Faulty values, while not inherently erroneous, stem from an unreliable data production process. Outliers may arise at the beginning and moreover may not be inherently erroneous; however, they manifest as data points deemed intuitively as 'anomalies', 'irregularities', 'deviations' or 'inconsistent observations' relative to a reference set of knowledge (Muhr *et al.* 2023). This ambiguity in defining outliers can be rationalised in the proposition that a continuum exists between 'good' values and outliers. Such a concept was notably explored in the context of health and medicine by Canguilhem (1966) and delves into the differentiation between normality and pathology. Consequently, identifying outliers proves to be a challenging and subjective endeavour. Outliers are sometimes labelled as 'abnormal', i.e. deviating from a norm. Yet this norm might not necessarily adhere to a normal distribution.

Clemens-Meyer *et al.* (2021) also underscored the subjectivity, hence variability, in data validation. This aspect is contingent upon the specific objectives and intended utilisations of the data, including system description and behaviour prediction. In the context of the five criteria listed above, evaluating data quality entails:

- **Defining key factors:** This involves the development of indicators, in discerning whether they should be quantitative or qualitative, along with the establishment of pertinent metrics or categories.
- **Defining a categorisation method:** This involves the determination of thresholds or reference values against which observations can be compared.

A deficiency exists in the auditability surrounding data processing, with data validation often being either intuitively performed or omitted in the reports due to its habitual integration into data production routines. Prevalent, yet infrequently documented, methods for handling missing data include interpolation, removal, imputation and replacement, as reported

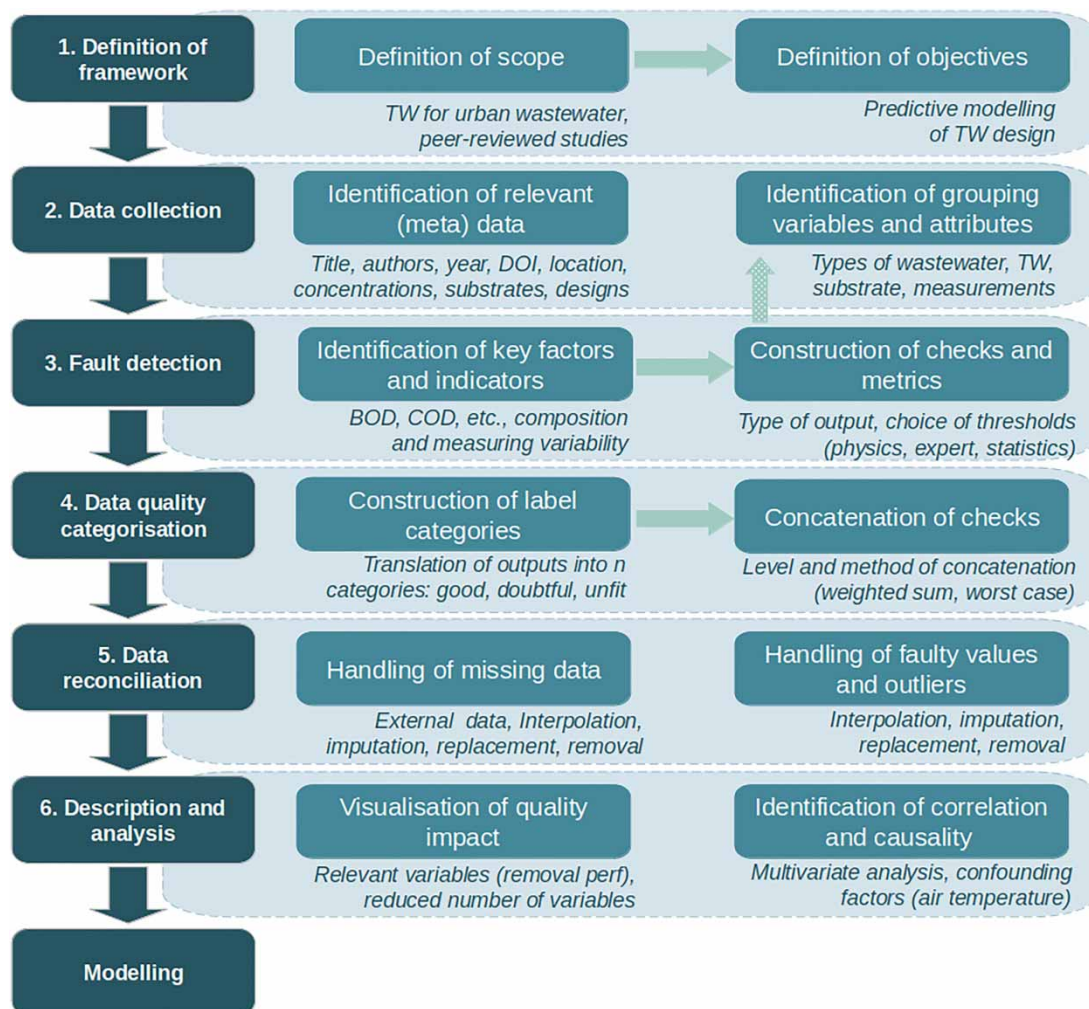
in [Zhu \*et al.\* \(2023\)](#). Similarly, data reconciliation is another common but underreported handling of experimental data and process data ([Rieger \*et al.\* 2010](#); [Cunha \*et al.\* 2021](#); [Ramasamy \*et al.\* 2021](#)). It consists of detecting and correcting for random noise and errors; data reconciliation typically occurs after fault detection or data categorisation.

## Objectives

The primary objective of this study is to establish a methodology for data validation in order to ensure reliable data for subsequent modelling efforts. We intend herein to elucidate the implicit decisions inherent in the production and management of data, thereby fostering awareness of research practices. Concurrently, the methodology is defined jointly with the development of a workflow customised for TW. Our secondary objective is to identify current expert knowledge, key factors, indicators and criteria relevant to TW specifically for the quality assessment of a dataset. Last, we provide a brief description of the relationships identified between variables, to be mobilised for modelling purposes.

## METHODOLOGY

[Rieger \(2012\)](#) and [Clemens-Meyer \*et al.\* \(2021\)](#) have devised methodologies tailored for the validation of time-series data from a single site. These methodologies mobilise historical data and a cumulative understanding of the system being monitored. Consequently, we have adapted these methodologies to validate a dataset sourced from the literature, in compiling information from diverse sources and sites. A methodology has been defined in six steps, each with two actions ([Figure 1](#)); it is



**Figure 1** | Data validation methodology in six steps and corresponding tasks to perform for each step (with examples in italics). Loops can be envisaged between the various process tasks.



open to both iterations and loops. Our emphasis has been placed in assessing the dataset against the criteria of plausibility, consistency and auditability.

## Step 1: Definition of framework

### Substep a: Definition of scope

This study focuses on horizontal flow treatment wetlands (HF wetlands) and vertical flow treatment wetlands (VF wetlands) for urban/domestic wastewater, at the global scale. The scope of this study has been restricted to experimental data (from pilot to full scale) published in peer-reviewed scientific journals.

### Substep b: Definition of objectives

The sole objective is to obtain reliable data for future works: identification of explanatory variables for treatment processes and performance, plus the definition of a predictive data-driven model for TW design.

## Step 2: Data collection

A bibliographic search was performed on SCOPUS using the query: TITLE-ABS-KEY ((horizontal AND flow OR hsf OR treatment AND wetland OR constructed AND wetland) AND (wastewater OR wastewater AND treatment)) AND (PUBLICATION YEAR > 2000). Both HF and VF wetlands were reported when used in multi-stages or jointly studied for comparative purposes. We then added papers selected by another study (Acuña *et al.* 2023) from WOS.

The papers were selected by means of a three-stage process: coarsely at first, based on the title/keywords/abstract; then more finely based on a screening of specific terms; and last, in detail by a thorough reading of the papers. We used R language (Team 2023) to screen all the pdf files, and the data were processed using the pdftools (Ooms 2023) and tm packages (Feinerer *et al.* 2008).

### Substep a: Identification of relevant (meta) data

Specific terms were sought within the content of the pdf files (excluding the references section). We used the following term-related criteria to pre-select the articles for a full reading:

- At least one occurrence of a dimensional term from among the following: surface area, length, width and depth.
- At least four occurrences of a pollutant term among the following: BOD, COD, NH<sub>4</sub>, NH<sub>4</sub>-N, NH<sub>3</sub>, NH<sub>3</sub>-N, NO<sub>3</sub>, NO<sub>3</sub>-N, NO<sub>2</sub>, N<sub>2</sub>O, TKN, TN, PO<sub>4</sub>, PO<sub>4</sub>-P, TP and TSS.

Raw data were reported within a relational database structured according to the categories outlined in Table 1, with an emphasis on meta-data showing the categories 'source' and 'treatment plant'. We defined a series of mandatory and optional fields, in targeting elements anticipated to be correlated with constituent removal mechanisms and performance (Supplementary Table S3).

**Table 1** | Categories and mandatory fields used to determine whether to keep or reject the articles

Category (or Table)	Mandatory fields	Additional fields
Source: on the publication	Type of document, title, authors, publisher, year, DOI	–
Treatment plant: on each TW treatment chain	Water type (type of influent), country	City, scale, number of stages
NBS TU: on each HF or VF stage in series in the TW treatment chain	Type, chain position, total surface area (m <sup>2</sup> ), depth, length or width (m), no. of parallel beds	Vegetation
Subsurface properties: on water saturation	–	Saturation
Layer properties: characteristics of the substrate of each step	–	Type of substrate, porosity, granulometry, d10, D60
Operational conditions: on hydraulics	Inflow rate (m <sup>3</sup> /d)	Hydraulic retention time
Sampling conditions: characteristics of the wastewater	–	Air/water temperature, conductivity, pH
Concentrations: on at least 1 compound among BOD, COD, NH <sub>4</sub> (-N), TN, TP, PO <sub>4</sub> (-P), TSS	Mean inflow and outflow values (concentration) (mg/L)	SD/min/max/median concentrations

For the dataset on TW, we defined an additional set of variables that had been computed from others, such as pollutant loads, and hydraulic and organic loading rates. We also provided units using the R package *units* from [Pebesma et al. \(2016\)](#), as well as data for the locations, consisting of the average temperature of the four coldest months, the average precipitation sourced from [Karger et al. \(2017\)](#) and the type of climate from [Chen & Chen \(2013\)](#). The correspondence between datasets is achieved through the geographic coordinates of the locations (city or country). We ran the *Nominatim* tool in the *GeoPy* package from Python, along with the *terra* package from R.

For good auditability, previous versions of the current dataset were saved as follows:

- Raw (meta) data were stored in a database hosted by ICRA.
- The final datasets were provided on HF and VF wetlands, along with all check functions developed in the R package *twwhybridmodel*, as obtained before performing data validation.

The following sections have grouped the data into HF as primary treatment, HF as secondary treatment, VF as primary treatment and VF as secondary treatment. In the presence of a multi-stage TW, we considered the first stage only.

### Substep b: Conventions of equivalence, identification of relevant grouping variables and attributes

The definition of the scope, collection and analysis of the data has relied on a definition of the entities to be studied, by means of determining the attributes that establish the similarity and discrimination between them, i.e. the categories of equivalence ([Desrosières 1992, 2006](#)). Categorisation consists of both a taxonomic operation (defining common traits) and an identification operation (fitting observations to the categories), the latter of which serves as the basis for data validation. We have adopted the following conventions:

TWs were differentiated, from an engineering perspective, into vertical and horizontal types based on the direction of water flow within the medium ([Fonder & Headley 2013](#)). We retained this classification, i.e. HF and VF wetlands which differ by the flow rate direction, substrate granulometry and number of parallel beds alternately fed. The scope was limited to the total surface areas greater than 1 m<sup>2</sup>, encompassing various scales: pilot, household, pilot-household and WWTP. A series of multiple TWs was sometimes referred to as a single TW system. We, however, considered each stage individually, in referencing its specific position within the treatment chain.

*Wastewater taxonomic operation.* Wastewater types were differentiated based on the source or treatment step where the wastewater originated. Water categories were aggregated under the following names:

- Raw wastewater: Any type of wastewater from among the household, domestic, residential, pilot or WWTP scales, without a treatment step.
- Primary treated wastewater: Any type of raw wastewater that had also undergone an initial treatment stage from among any type of 'primary treatment' or 'pretreatment'. Primary treatment includes settling, septic tank, Imhoff tank, upflow anaerobic filter and sludge blanket and anaerobic pond.
- Secondary treated wastewater: No discrimination was made among the various secondary treatments.

*Wastewater identification operation.* The conformity of wastewater composition to the type it should belong to is assessed by checking the concentrations and typical concentration ratios between pollutants. We compiled and defined as expert knowledge the intervals of concentrations of the different pollutants and their ratio intervals ([Tables 2 and 3](#)) defined by the following authors: [Canler & Perret \(2007\)](#) for 'secondary wastewater' (wastewater treated by activated sludge); [Henze et al. \(2008\)](#) for urban wastewater; [Mercoiret \(2009\)](#) with 10,000 observations from WWTP across France; and [Hauduc \(2011\)](#) with 10 observations from WWTP throughout the world.

TW substrates were grouped into more global categories based on the main layer composition:

- Sand: any type of sand
- Gravel: any type of gravel, limestone
- Granules: ceramsite, Leca
- Other minerals: tuff, slag, tezontle
- Other organics: sphagnum, soil, mulch
- Unknown: when the substrate was not mentioned.

**Table 2** | Intervals for typical concentrations in raw urban wastewater (from Henze *et al.* (2008) and Mercoiret (2009))

Pollutant	Raw wastewater concentrations (mg/L)				Secondary treated concentrations (mg/L)	
	Min	Average	Median	Max	Min	Max
BOD	39.0	265	250	570.0	10	15
COD	122.0	645.7	604	1,341.0	60	70
NH <sub>4</sub> -N	12.0	54.9	55	98.3	–	–
TKN	14.1	67.3	67	123.1	5	6
TP	2.0	9.4	9	18.4	1.2	1.5
PO <sub>4</sub> -P	6.0	–	–	10.0	0.2	0.3
TSS	53.0	288	240	696.0	15	20

**Table 3** | Intervals for typical ratios in raw urban wastewater (from Henze *et al.* (2008), Mercoiret (2009) and Hauduc (2011))

Ratio	Raw wastewater				Primary treated				Secondary treated			
	Min	Mean	Max	Median	Min	Mean	Max	Median	Min	Mean	Max	Median
COD:BOD	1.800	2.6	3.90	2.5	0.5	1.874	3	1.9	4	–	7	–
COD:TN	6.670	10.52	20.00	10.99	2.78	7.46	20	8.33	–	–	–	–
COD:TP	38.460	62.5	112.36	66.67	16.67	43.48	100	43.48	40	–	58.33	–
COD:TSS	1.270	2.17	4.35	2.38	1.79	2.63	5.56	2.5	3	–	4.66	–
BOD:TN	3.000	–	8.00	–	–	–	–	–	–	–	–	–
BOD:TP	12.600	28.5	47.00	26.8	–	–	–	–	6.66	–	12.5	–
NH <sub>4</sub> -N:TKN	0.500	0.74	0.97	0.74	0.43	0.755	0.9	0.75	–	–	–	–
BOD:TKN	1.900	3.88	6.50	3.7	–	–	–	–	–	–	–	–
TKN:COD	0.063	0.12	0.18	0.11	–	–	–	–	–	–	–	–
BOD:NH <sub>4</sub> -N	5.300	–	7.69	–	–	–	–	–	–	–	–	–
PO <sub>4</sub> -P:TP	0.390	0.603	0.80	0.6	0.5	0.741	0.9	0.75	–	–	–	–

Variables were grouped according to the common category of reference:

- Scale: length, inflow rate, depth, HLR, HRT, total surface area
- Design: ‘nominal’ surface area, ‘nominal’ cross-section
- Climate: air temperature, influent water temperature, precipitation, altitude
- Water: influent conductivity, influent pH
- Substrate: porosity, minimum granulometry, delta granulometry
- Inflow concentrations: expressed in mean concentration of the six pollutants
- Removal performances: for all six pollutants.

Most studies cover a long period and only report the average values of measured concentrations. We will use the term ‘observation’ herein to indicate the average of concentration measurements reported for a given pollutant, while ‘sampling campaign’ refers to the period covered in a given study. No discrimination was made between grab (punctual) and composite (averaged) samples. When provided, the standard deviation, median, minimum and maximum values were all collected. The data were homogenised by means of converting pollutant concentrations to their N and P equivalents whenever applicable (Supplementary Table S1).

### Step 3: Fault detection

We have defined checks for data reliability, in questioning the extent to which the observations comply with certain criteria; next, quality weights, labels and flags were assigned to the observations.

#### Substep a: Identification of key factors and indicators

We identified the following key factors, which should be checked for quality:

- Type of influent wastewater. It seems relevant to check the adequacy between the taxonomic operation (type of wastewater indicated) and identification operation (composition). We applied the typical concentration and ratio ranges as indicators.
- Concentration measurements. When measuring the concentrations of multiple organic carbonaceous, nitrogen or phosphorous components, inconsistencies may arise across measurement method outputs. We defined a set of physical relationships such that their concentrations should be respected as indicators.
- Water composition. The case studies are examined together as a global population of regionally and scale-wise distinct and independent subpopulations. Their features are assumed to follow similar trends. Concentration was chosen as an indicator.

#### Substep b: Construction of checks and metrics

We decided to base our distinction of outliers and faults from good values with the three following checks, in opting for a continuous representation of values. Two types of outputs were derived: the probability of obtaining such a value and an associated label for better visualisation.

*Accuracy check – influent water type and concentration.* Two checks were defined. For each pollutant and ratio of pollutants, values were compared with the typical concentration and ratio ranges defined by the expert knowledge defined in Tables 2 and 3. The closeness to expected ‘true’ values could therefore be assessed for each type of wastewater, with outliers lying furthest from the reference values.

- To fit a law, we assumed the expert min–max intervals comprise 90% of the distribution.
- The probability thresholds were defined for quality labelling such that the tolerant range comprises 95% of the distribution [0.025, 0.975], while the stringent range comprises 90% of the distribution [0.05, 0.95].

#### Concentrations

- We fitted a lognormal distribution  $\log N(\mu, \sigma^2)$  to the expert data (Table 2) for each pollutant, in order to represent the concentration distribution.
- We optimised  $\sigma$  such that  $P[C < \min] \sim 0.05$  and  $1 - P[C < \max] \sim 0.05$ , i.e. using  $R$ , the value minimising the distance:  $\text{abs}(\text{plnorm}(\min, \mu, \sigma) - 0.05) + \text{abs}(1 - \text{plnorm}(\max, \mu, \sigma) - 0.05)$ , where  $\mu = \log(\text{median})$ , else  $\log(\text{mean}) - (\sigma^2/2)$ , or else if none is available,  $\log((\min + \max)/2)$ .
- We computed the probability of obtaining each observation value or higher according to the fitted distribution.
- We labelled cases as ‘unlikely’ in the event of non-compliance with the tolerant range, or else ‘unusual’ in the event of non-compliance with the stringent range. The output was named ‘Quality concentration’.

#### Ratios

- A normal distribution  $N(\mu, \sigma^2)$  was fitted to the expert data (Table 3) for each ratio, in order to represent the ratio distribution.
- We optimised  $\sigma$  such that  $P[\text{ratio} < \min] \sim 0.05$  and  $1 - P[\text{ratio} < \max] \sim 0.05$ , i.e. using  $R$ , the value minimising the distance:  $\text{abs}(\text{pnorm}(\min, \mu, \sigma) - 0.05) + \text{abs}(1 - \text{pnorm}(\max, \mu, \sigma) - 0.05)$ , where  $\mu = \text{mean}$ , else  $\text{median}$ , or else if none is available,  $(\min + \max)/2$ .
- We computed the probability of obtaining the ratio value or higher according to the fitted distribution.
- For each pollutant, from the ratios specifically applicable to it, we computed the proportion of checks complying with both the tolerant and stringent ranges.



- We labelled cases as ‘unlikely’ if more than half of the tolerant checks or over 0.8 of the stringent checks were not validated, or else ‘unusual’ if over 0.2 of the tolerant checks or more than half of the stringent checks were unvalidated. The output was named ‘Quality water type’.

*Consistency check – quality measure (in/out).* We checked that the measured concentrations satisfied the following physical relationships (inequalities). We then weakened the checks to simpler ones (by removing pollutants, replacing TKN by TN) to keep operating the checks in case not all pollutants were being measured:

- Organic carbonaceous components:  $[\text{COD}] > [\text{BOD}]$
- Reduced nitrogen components:  $[\text{TKN}] > [\text{NH}_4\text{-N}] + [\text{NH}_3\text{-N}]$  and  $[\text{TN}] > [\text{NH}_4\text{-N}] + [\text{NH}_3\text{-N}]$ ,  $[\text{TN}] > [\text{NH}_4\text{-N}]$ ,  $[\text{TN}] > [\text{NH}_3\text{-N}]$
- Oxidised nitrogen components:  $[\text{TN}] \sim [\text{TKN}] + [\text{NO}_3\text{-N}] + [\text{NO}_2\text{-N}]$  and  $[\text{TN}] > [\text{NO}_3\text{-N}] + [\text{NO}_2\text{-N}]$ ,  $[\text{TN}] > [\text{NO}_3\text{-N}]$ ,  $[\text{TN}] > [\text{NO}_2\text{-N}]$
- Phosphorus components:  $[\text{TP}] > [\text{PO}_4\text{-P}]$ .

To fit a law, we considered a one-sided check identifying when  $C_{\text{tot}} < \Sigma C_i$ , and assumed that the interval of differences  $[0, \text{gap} = 0.1C_{\text{tot}})$  comprised 90% of unverified inequalities.

We defined probability thresholds for quality labelling such that the tolerant threshold comprises 95% of the values, and the stringent threshold comprises 90% of the values.

- We fitted an exponential distribution  $\text{Exp}(\lambda)$  for each pollutant  $C_{\text{tot}}$  to represent the distribution of unverified inequalities.
- We optimised  $\lambda$  such that  $P[\text{difference} < \text{gap}] \sim 0.9$ , i.e. using  $R$ , the value minimising the distance:  $\text{abs}(\text{pexp}(\text{gap}, \lambda) - 0.9)$ .
- We computed the probability of obtaining each difference or higher according to the fitted distribution. The verified inequalities reach  $P = 1$ .
- For each component, from the inequalities specifically applicable to it, we computed the number of checks complying with both the tolerant and stringent ranges.
- We labelled cases as ‘unlikely’ if at least one tolerant check was not validated and as ‘unusual’ if at least one stringent check was not validated. These instances may serve as evidence for a faulty measurement method or choice of devices. The outputs were named ‘Quality measure in’ and ‘Quality measure out’.

*Plausibility check – Chauvenet (in/out).* It was statistically assessed whether a given concentration value, taken from the set of all inlet or outlet concentrations, is likely to be spurious. This check is similar to the accuracy check of concentration, with the difference being that external expert knowledge is not introduced. Notably, we used this check to compensate for the absence of expert data in primary treated wastewater.

We employed the Chauvenet criterion, as defined for an assumed normal distribution. This assumption could not be adopted for chemical concentrations in wastewater. After plotting the distribution of cases for each constituent, it was decided to fit a lognormal distribution  $\log N(\mu, \sigma^2)$  to the data (Oliveira *et al.* 2012; Von Sperling *et al.* 2020). Hence, for a given component at concentration  $C_i$  in the dataset, the associated probability  $P_{C_i}$  equals  $P_{C_i} = 1 - P_{\log N}[C < C_i]$ .

Unusual values were defined only on the right tail of the distributions, with:

- a stringent threshold at 0.1 probability of occurrence and
- a tolerant threshold at 0.05 probability of occurrence.

We labelled cases as ‘unusual’ if their probability  $P_{C_i}$  was below 0.1 and ‘unlikely’ if their probability was less than 0.05. The outputs were named ‘Compliance Chauvenet’ and ‘Compliance Chauvenet out’.

We additionally applied this check on the operational condition scale to assess the homogeneity of surface HLR values, fitting a lognormal distribution to the data. The output was named ‘Compliance Chauvenet HLR surface’.

## Step 4: Data quality categorisation and concatenation

### Substep a: Construction of label categories

For the quality flag, the number of labels was held to three, with a terminology referring to the prospect of modelling found in Clemens-Meyer *et al.* (2021), plus a fourth one when the checks could not be applied:

- ‘good’ (‘fit for use’) when observations complied with stringent checks, else
- ‘doubtful’ (‘questionable’) when observations complied with tolerant checks, else
- ‘unsuitable’ (‘unfit for use’) and
- ‘unchecked’ when the checks could not be applied.

#### Substep b: Concatenation of checks

- Per each observation: the outcomes of the different checks were concatenated into a quality flag by taking the worst outcome per observation: ‘unlikely’ yielding ‘unsuitable’, ‘unusual’ yielding ‘doubtful’ and ‘likely’ yielding ‘good’. The ‘unchecked’ label was not considered in the concatenation.
- Per each sampling campaign: the flags of the various observations were further concatenated into a single flag per sampling campaign. We computed the proportion of ‘unsuitable’ and ‘doubtful’ flags and labelled the sampling campaign as ‘unsuitable’ if more than half the flags were ‘unsuitable’ or over 0.8 of the flags were ‘doubtful’, else ‘doubtful’ if over 0.2 of the flags were ‘unsuitable’ or more than half the flags were ‘doubtful’.

### Step 5: Data reconciliation

Data reconciliation, in particular the handling of missing data, constitutes an optional step in data processing.

#### Substep a: Handling of missing data – data imputation

The missing data was handled for purposes of clustering and multivariate analysis, and not for models since such handling steps lead to higher artificial correlations between variables. With a small, non-time-series and heterogeneous dataset, we considered imputing missing data rather than other methods (interpolation, removal, replacement), which could lead to biased estimates and underestimated uncertainty. We observed data to be missing at random (MAR), given that missing data is not randomly distributed among variables, but randomly distributed for each variable across the case studies. More specifically, we considered no explanatory factor accounts for a missing data occurrence on a given variable, and we expect the full imputed dataset to follow similar observed distributions. We used the predictive mean matching method from the MICE package (Buuren & Groothuis-Oudshoorn 2011; Buuren 2018). This method randomly draws a value from the distribution of observed values for each missing data, as determined by an adjustable set of predictor variables (Figure S5). We generated one imputed dataset with additional constraints, thus ensuring that plausible values are imputed.

- For clustering: We considered a subset of the joint datasets of HF and VF wetlands, as primary and secondary treatment. It is composed of the inflow concentrations among BOD, COD, NH<sub>4</sub>-N, TN, TP and TSS. We imputed the non-measured inflow concentrations per sampling campaign, and iterated the process on the imputed values that yielded ‘unsuitable’ labels during the consistency/accuracy check.
- For multivariate analyses: We imputed the full datasets separately for HF and VF wetlands, as secondary treatment. The predictive variables were chosen to either precede chronologically or be considered constant with respect to the predicted variables (Supplementary Figure S5).

#### Substep b: Handling of faulty values and outliers

We assessed the information, proportion and leverage of the data labelled as ‘unsuitable’. For better visualisation, the ‘unsuitable’ values were to be removed from the dataset. For modelling purposes, however, the very low values of quality weights would avoid removing them.

### Step 6. Data description and analysis

Our dataset was described and compared to the reference values from Cross *et al.* (2021). We then assessed the adequacy between the types of water and clusters obtained with imputed data by performing hierarchical clustering, from the factoextra package (Kassambara & Mundt 2020). We compared the use of a Euclidean metric (usual) and Pearson metric (similarity between observations) on the influent wastewaters.

We performed a multivariate exploratory analysis of the dataset in order to identify pairwise correlations and potential causality. To ease the data visualisation and interpretation tasks, we relied on two methods from the FactoMineR package (Lê *et al.* 2008):

- MFA (Multiple factor analysis): We grouped the variables from the correlation plots under a common reference (climate, design, inflow concentrations, scale, water). MFA weights the variables within each group.
- PCA (Principal component analysis): We used the variables from the correlation plots, in removing one of the variables strongly correlated pairwise so as to avoid multicollinearity. PCA provides, for numerical variables, the amount of variance explained.

We put removal performance and flag as supplementary variables.

## RESULTS AND DISCUSSION

### Steps 1 and 2: Definition of framework and data collection

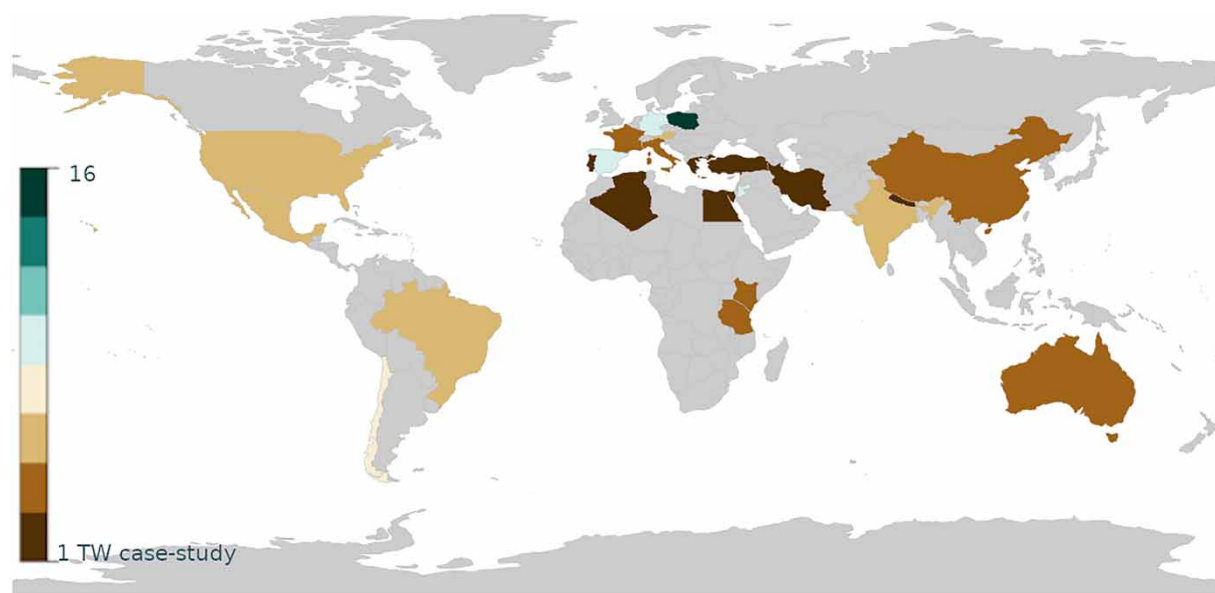
For the data collection part, we selected the articles over several steps and obtained the following numbers:

- 507 articles from the initial query (title, keywords, abstract).
- 380 articles from the automatic screening of their abstract, depending on the terms searched. The number of publications substantially increases over the period 2015–2019: the median year for the search period (1995–2022) is 2017 (general description in Supplementary Figure S1).
- 291 articles from an automatic screening of their pdf.
- 68 articles after a full reading. The final selected studies were few in number compared to the original number of case studies, due to overly limited information on the designs and contexts. Let's recall that the goal herein is to build a dataset for modelling purposes and not just for reporting on removal performance. Some studies concerned both types of TW; therefore, we ultimately used 54 papers to build the dataset on HF wetlands and 31 papers for that on VF wetlands.

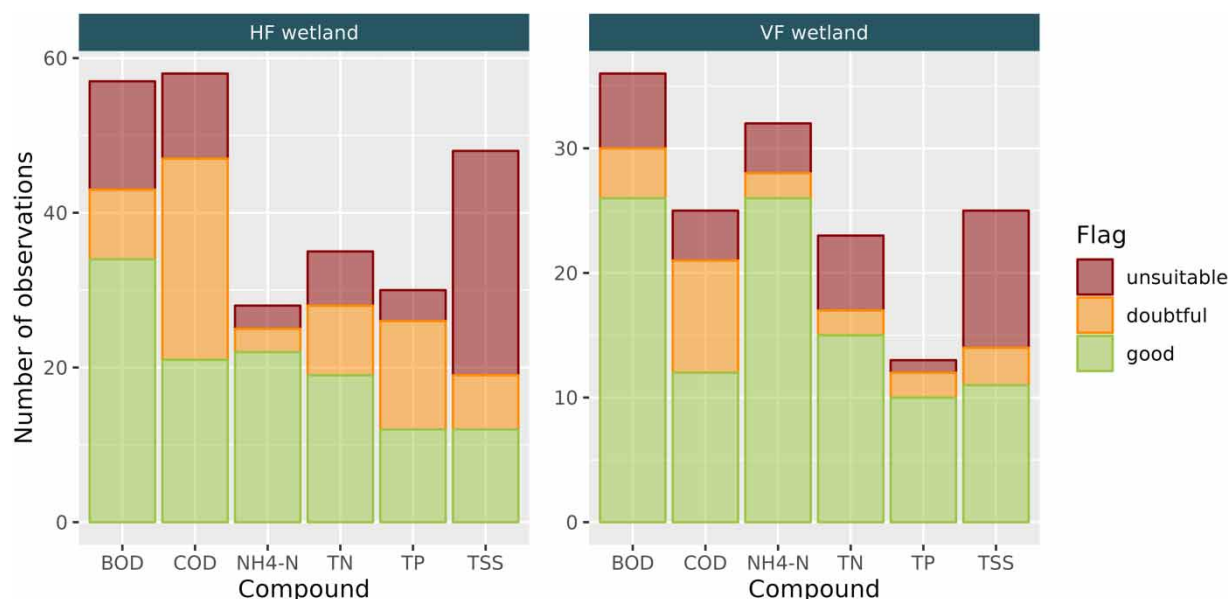
After defining the equivalence categories, both of the stored datasets contained a majority of TW used for primary treated wastewater: 62 HF wetlands (with 68 sampling conditions) and 33 VF wetlands (with 37 conditions) (Supplementary Table S4); and 12 HF wetlands (30 samplings) and 10 VF wetlands (13 samplings) pertained to the treatment of raw wastewater (Supplementary Figure S2).

Despite a substantial reduction in the initial number of case studies, we considered the distribution of study locations across the globe to provide a representative set of the different climates and urban wastewater compositions (Figure 2).

BOD was reported in 87 and 82% of sampling reports for HF and VF wetlands, respectively, while COD came in at 86 and 76%,  $\text{NH}_4\text{-N}$  at 49 and 91%, TN at 56 and 69%, TP at 43 and 31% and TSS at 74 and 76% (Figure 3).  $\text{NH}_4\text{-N}$  is more often reported for VF wetlands than HF wetlands given their expectation to achieve high removals; moreover, TP is



**Figure 2** | Number of treatment wetland steps across the world, HF and VF wetlands together.



**Figure 3** | Number of observations per constituent for HF as secondary treatment and VF wetlands, with quality flags.

overall less frequently measured since TWs are not designed for removal, and regulations only consider thresholds for eutrophication-sensitive sites. Nonetheless, from the perspective of treated wastewater reuse, it may be found useful to measure TP.

### Categories of equivalence

The expert knowledge gathered for the identification operation of wastewater type presents overlapping criteria between the defined categories of equivalence (Tables 2 and 3). These boundaries blurred for the concentration and ratio criteria indicate a potential use in detecting unrealistic values yet remain insufficient in identifying a type of wastewater *a posteriori*. We applied bottom-up hierarchical clustering, with  $k = 2$  clusters, on the sampling campaign dataset composed of influent raw and primary treated concentrations (Figure 4). The taxonomic operation also presents an overlap between raw and primary treated wastewaters (fourth graph). The method with the Euclidean metric primarily discriminated between low and high concentration values, whereas the method with the Pearson metric differentiated the dataset into clusters with greater similarity to the distribution between raw and primary treated wastewater, thus suggesting that the relations between pollutants participate in discriminating the types of wastewater. The algorithm attributed 30 out of 42 raw wastewater sampling campaigns to Cluster 1, but just 49 out of 103 primary treated wastewater campaigns to Cluster 2. The taxonomic and identification definitions of the types of wastewater only allow for weak discrimination.

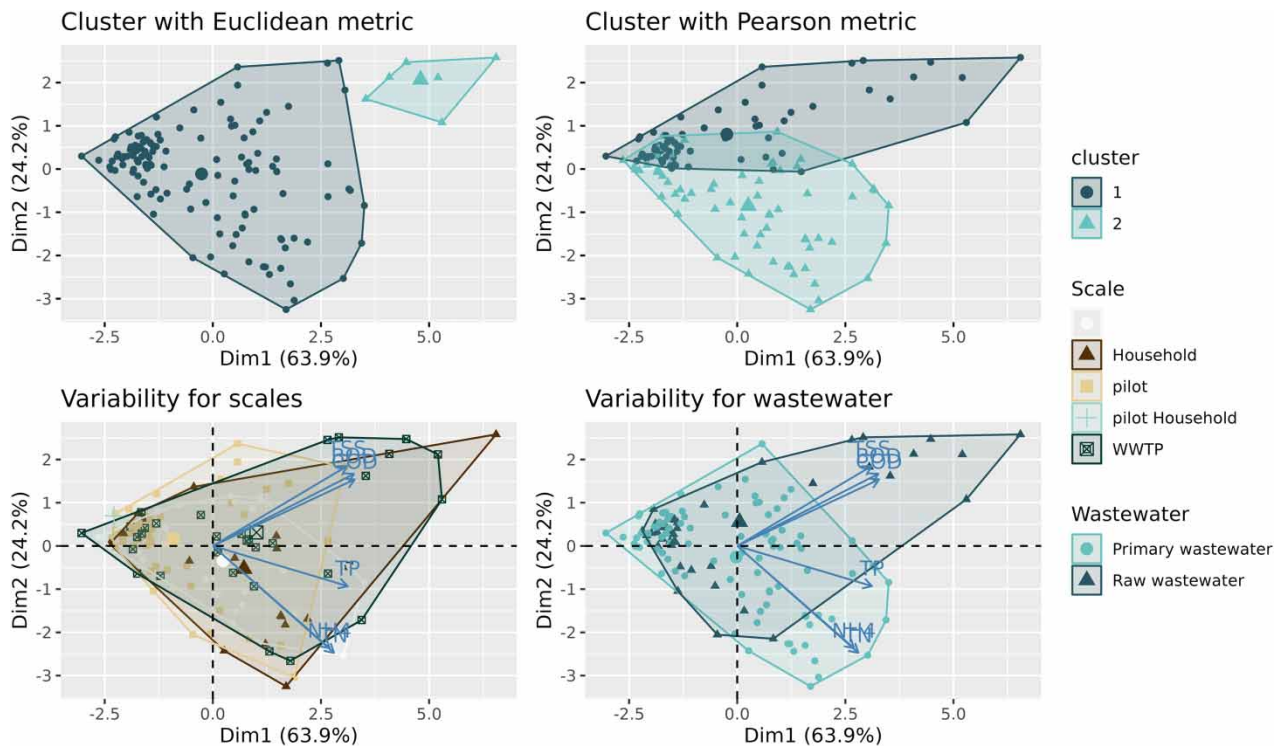
The categories of equivalence for the scale do not coincide with the clusters and moreover overlap between one another, revealing that the wastewater compositions did not vary across scales. The pilot scale, however, only displayed lower concentrations.

### Step 3: Fault detection

#### Substep a: Key factors and indicators

The main characteristic of the constituent concentrations reported in the case studies was the mean, followed by the standard deviation, and seldom the median, minimum or maximum values. In most instances, the mean was higher than the median value, as would be expected since Oliveira *et al.* (2012) observed that inflow and outflow concentrations for multiple treatment systems follow a lognormal distribution, although TWs were not studied.

For each constituent, we collected more than 60 observations, except for TP, which consisted of 44 observations. This volume of data was deemed sufficient to model a distribution based on the mean concentrations observed in influent primary treated wastewater. Our analysis revealed a positively skewed distribution, an aspect that could be partially attributed to the



**Figure 4** | Clusters obtained and real labels for one imputed concentration dataset for influent raw and primary treated wastewater samplings, with  $k = 2$  clusters. The Euclidean and Pearson metrics were compared to define the distances for the clustering method, both with and without additional variables corresponding to the ratios. The concentrations have been scaled.

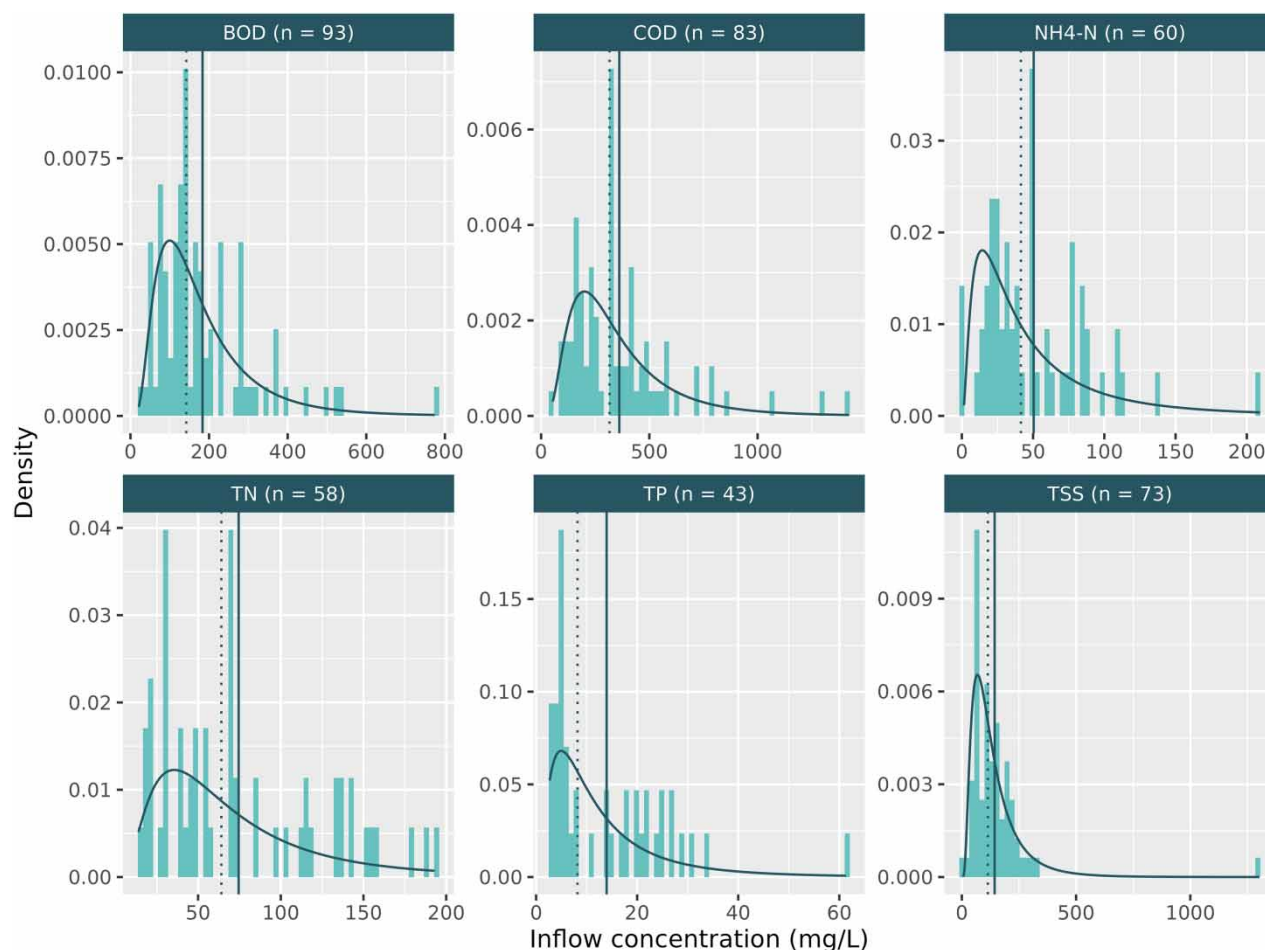
inherent constraint that concentrations must possess positive values. Consequently, this constraint inherently results in a scenario whereby the median value is consistently smaller than the mean concentration (Figure 5 and further details in Supplementary Figures S3 and S4). This mean–median difference suggests that over 50% of the time, the TWs function on concentrations below the mean value. The Shapiro–Wilk test led to the conclusion that the resulting dataset is unlikely to be observed if the parent population is normal with an alpha risk of 5% for all constituents ( $H_0$ : ‘The distribution is not significantly different from a normal distribution’). This observation highlights that the use of statistical tools relying on symmetry and normality hypotheses, which are non-compliant with the dataset, reduces their reliability. We decided to adapt the plausibility checks by using lognormal distributions on concentrations derived from results on the positive skewness, output of the normality test and the domain of definition (positive values). Nevertheless, the limited number of observations does not serve to statistically base the study for raw wastewater. Yet we can infer a similar conclusion from the theoretical approach of the principle of proportional effects, as based on the central limit theorem (Supplementary Figure S3).

As a measure of the overall trend, use of the mean rather than the median or mode has predominated over time with a shift of paradigm, from the search for the ‘true’ or ‘right’ value to a more abstract and summary value. In practice, the mean is associated with a normal distribution, i.e. interpreted as the most usual or ‘normal’ value assigned to a variable. However, the ‘norm’ followed by a variable may not in fact be a normal distribution; it might be skewed, and the mean may not coincide with a real-world case. The mean remains interesting as a barycentre for non-normal distributions. When reporting a dataset summary, Von Sperling *et al.* (2020) recommends expressing the standard deviation in parentheses after the mean value, so as to avoid implying a symmetrical dispersion. We also suggest adding the skewness and the number of observations.

### Substep b: Checks and metrics

The studies do not always monitor the full set of constituents or pollutants of the same family; consequently, checks on measurement quality could not be performed in most cases (see checks in Figure 6 and Supplementary Figure S12 for the details per pollutant and label). We noted that primary TWs present a higher proportion of ‘unsuitable’ flags than secondary





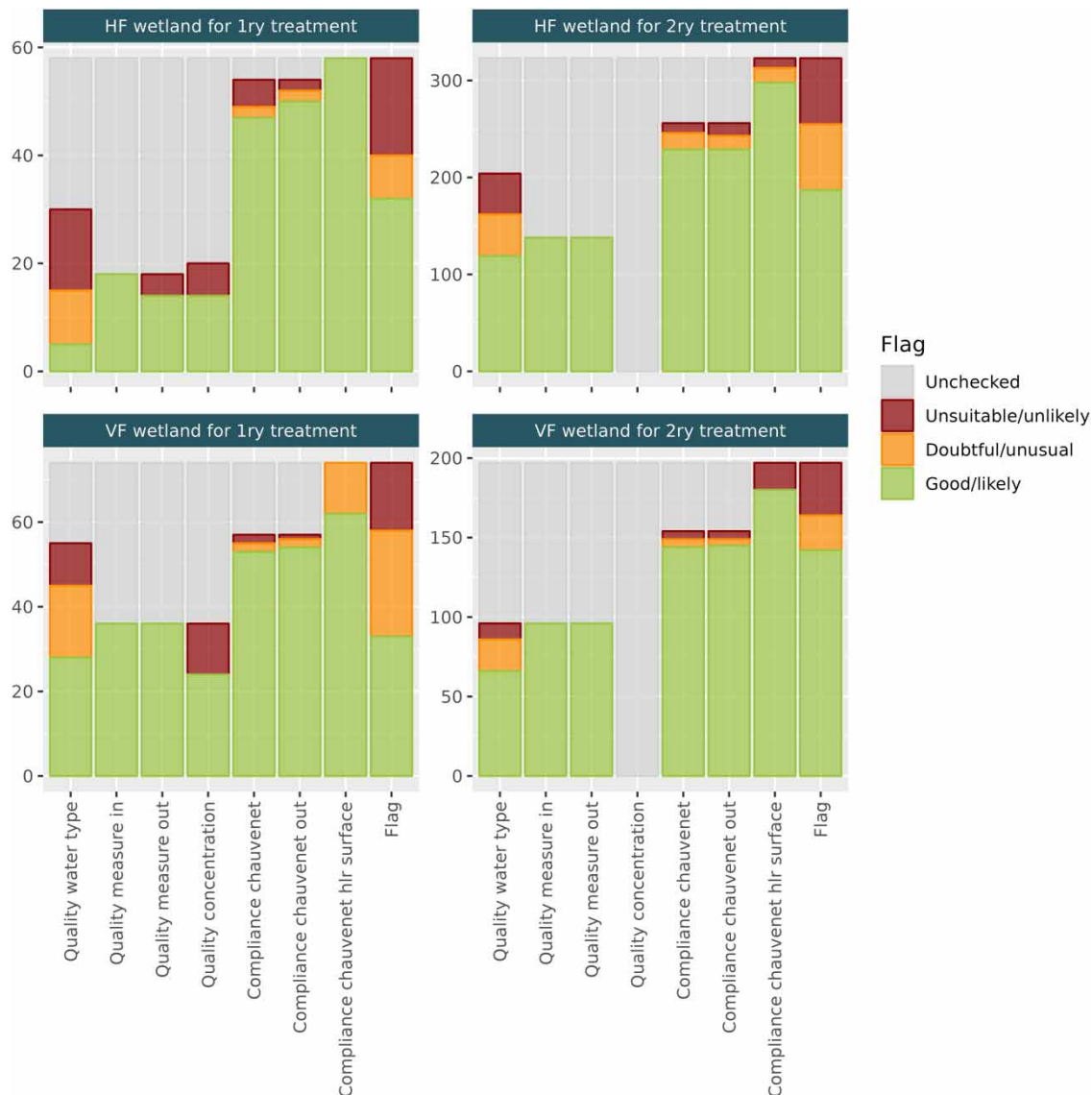
**Figure 5** | Distribution of mean constituent inflow concentrations in primary treated wastewater, with the mean value (solid vertical line), median value (dashed vertical line) and fitted lognormal distribution (dark blue curve).

TWs due to the possibility of performing an additional check on influent concentrations. For the check on water type, half of the ratios are ‘unsuitable’ for HF wetlands, while nearly a third of the ratios and concentrations are ‘unsuitable’ for VF wetlands. Since water uses and compositions as well as treatment technologies fluctuate across the world, the water categories (or their boundaries) established from a given reference may become obsolete, i.e. their use would be inappropriate to extrapolate to a different context.

The data validation methodology is linear. In practice, iterations and jumping back into the workflow have proven to be necessary for us to adjust the choice of equivalence categories or the definition of checks. We defined categories to facilitate the visualisation of outputs; however, the probabilities obtained for each check serve as a quantitative measure of trustworthiness that is easier to use as a weight for subsequent data quality-weighted modelling. This categorisation also avoids discarding too many or too few data points due to a dichotomous labelling.

*Expert knowledge.* Expert knowledge to use as reference was difficult to collect. Notably, the check for ranges of concentrations could not be performed on primary treated wastewater due to a lack of expert knowledge. Our accuracy and consistency checks also relied on tenuous assumptions regarding the choice of thresholds and distributions of concentrations, ratios and measurement gaps.

The expert knowledge derived mainly concerned European practices. The composition, volume and concentrations of wastewater may substantially vary depending on water usage. Water production indeed differs across regions: as an example,  $209.5 \text{ m}^3 \cdot \text{year}^{-1}$  per capita in North America vs.  $67.6$  in Latin America and the Caribbean (Jones *et al.* 2021). Therefore, the



**Figure 6** | Number of observations per quality label per check for both primary and HF as secondary treatment and VF wetlands.

typical values and ranges used to compare case studies may be irrelevant. In particular, some very high mean inflow concentrations (BOD, COD, TN) corresponded to a snowy, fully humid with warm summer climate (Dfb), and moreover were labelled 'doubtful' or 'unsuitable' during the plausibility check using the Chauvenet criterion. These might not be outliers but rather typical cases in the corresponding regions (India, Nepal and Poland), where lower water usage and volume lead to higher concentrations. The performance recorded in identifying 'outliers' for the checks based on expert knowledge thus increases with both dataset homogeneity and closeness to the scope of the expert knowledge. In other words, the power of our checks increases when performed on subpopulations. So, when applied to a context larger than the reference, the reliability of the checks decreases. A unique statistic applied over a large scope is not as efficient at detecting outliers.

As regards statistical methods, the Chauvenet criterion has led to labelling more observations as 'doubtful' and 'unsuitable', compared to the Z-score, which is often used for outlier detection (Supplementary Figure S11); however, we consider this difference to be negligible. We would suggest use of the Chauvenet criterion, which provides a probability (i.e. a 'quantification of uncertainty') that is more easily interpretable (aligning with [Muhr et al. \(2023\)](#)) and better grasps the continuity between typical and outlier values than the Z-score.

#### Step 4: Data categorisation

The categorisation process led to labelling as ‘unsuitable’ 33% of HF as primary treatment and 21% of HF as secondary treatment, along with 35% of VF as primary treatment and 19% of VF as secondary treatment, as shown in Figure 6. The final quality indicator, i.e. the flag, depends to a great extent on the elements and method used to concatenate. The data validation methodology resulted in potentially discarding up to a third of the observations (unsuitable, in red) for some of the constituents (Figure 3). The accuracy check on the type of water (ratios) gave rise to most of the ‘unsuitable’ flags, notably more than half for TSS in HF wetlands, which could be due to the lone tested ratio COD:TSS, with a narrow expert range, followed by the plausibility check using the Chauvenet criterion. Meanwhile, the consistency check on the measurements was hardly limiting at all (Figure 6 and, in detail, Supplementary Figure S12). For the type of water,  $\text{NH}_4\text{-N}$  was almost never checked, whereas for the measurements, the main relationship checked was  $\text{BOD} < \text{COD}$ .

At the observation level, we elected to concatenate with the most conservative label. It, however, does not result with a fine discrimination as two observations can obtain the same flag despite having different proportions of that ‘worst’ label. Other possibilities consist of taking the ‘mean’, ‘weighted sum’ or ‘median’ value of the checks. The weighted sum proves to be interesting whenever special indicators or key factors are to be prioritised.

#### Step 5: Data reconciliation

A substantial variability occurs between imputed datasets for variables with few observations, as regards the sampling conditions. The imputing method leads to higher correlations between certain variables, yet we considered this approach more realistic than others (Buuren 2018). In particular, the generation of multiple imputed datasets allows the handling of uncertainty and its inclusion in modelling. We did not modify or remove any faulty values or outliers since it had been emphasised that process-based corrections can also induce bias, as outlined in Villez (2017).

However, for modelling purposes, we opted to discard values considered to be isolated for the inflow rate and the surface area variables, i.e. when a gap exists in the order of magnitude between the values and the rest of the data. We excluded inflow rates and surface areas as key factors for quality assessment, although they could be mobilised for an accuracy check during the identification of the categories of scale. Providing a sufficiently large and homogeneous dataset, ‘outliers’ could be identified using the Chauvenet criterion. The outliers should be discarded since they may induce high leverage on the fitting of a statistical model and greatly modify its domain of definition.

#### Step 6: Dataset description and multivariate analysis

We observed that the hydraulic loading rate (HLR) and hydraulic retention time (HRT) were reported in various ways; most of the time, the value is divided by the surface area, thus suggesting (in the case of HF wetlands) its use as a proxy for the TW volume to adimensionalize the metric. For HF wetlands, we suggest reporting HLR computed for both the surface area and cross-sectional area of each TW, or in the case of multi-stage TW, of each first stage in order to offer an additional indication on the flow rate and propensity to clog. Pucher & Langergraber (2019) found HLR to be correlated with clogging in the case of VF wetlands.

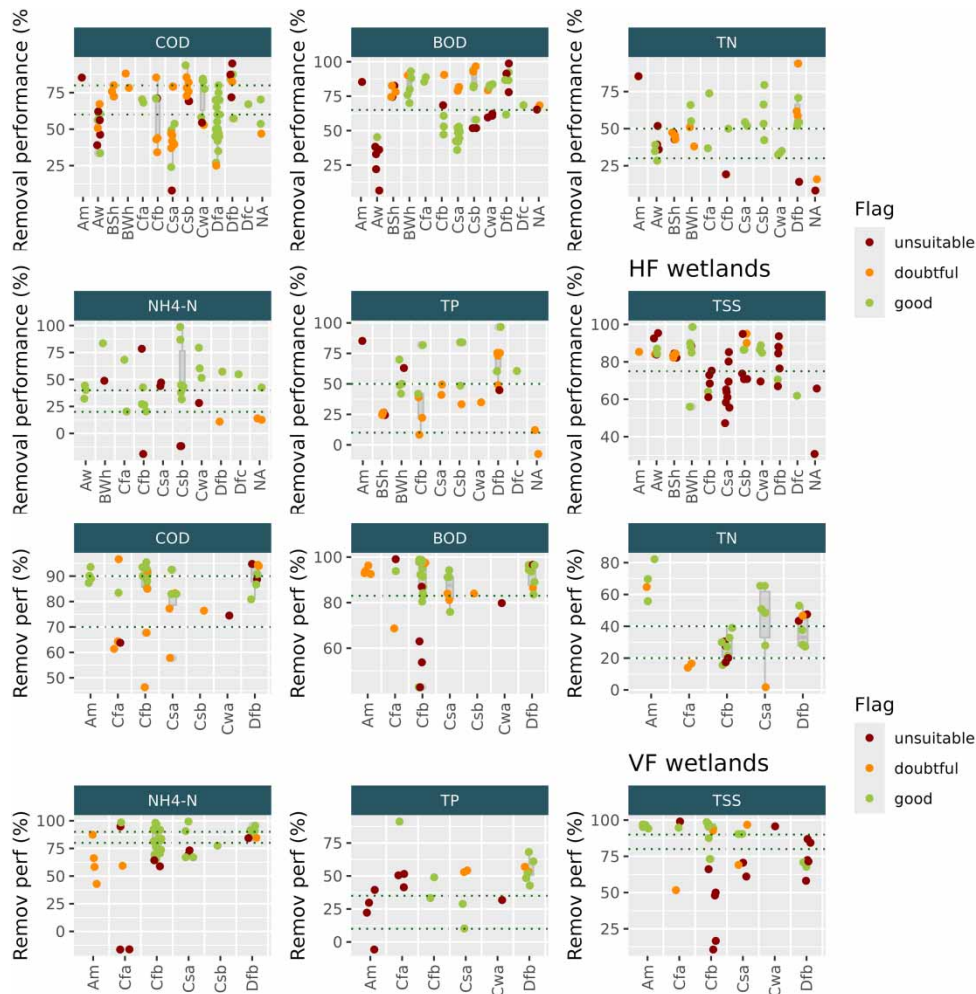
The surface areas were reported sometimes as that of the full treatment chain (comprising multiple TWs in series) rather than separately, for each stage. As a consequence, when recomputing HLR for operating first stage HF wetlands, nearly all HLR based on the surface area lie above the usual range reported in Cross *et al.* (2021):  $0.02\text{--}0.05 \text{ m}^3\cdot\text{m}^{-2}\cdot\text{d}^{-1}$ , with a median of  $0.10 \text{ m}^3\cdot\text{m}^{-2}\cdot\text{d}^{-1}$ . The observed median for the HLR based on the cross-sectional area equals  $1.36 \text{ m}^3\cdot\text{m}^{-2}\cdot\text{d}^{-1}$ .

The median OLR based on the surface area is 14.24 for BOD and 31.29 for COD; moreover, many observations lie above the typical operation conditions reported in Cross *et al.* (2021):  $8 \text{ g BOD m}^{-2}\cdot\text{d}^{-1}$  and  $< 20 \text{ g COD m}^{-2}\cdot\text{d}^{-1}$ . The observed median OLR for TSS is 7.22 which is slightly below the typical HF wetlands value of  $10 \text{ g TSS m}^{-2}\cdot\text{d}^{-1}$ . The observed median OLR based on the cross-sectional area equals 169.15 for BOD, which lies below the threshold of  $< 250 \text{ g BOD}\cdot\text{m}^{-2}\cdot\text{d}^{-1}$ . A few observations rose to 4–10 times the threshold.

Conversely, for VF wetlands, nearly all HLR based on the surface area lie under the typical value of  $0.1 \text{ m}^3\cdot\text{m}^{-2}\cdot\text{d}^{-1}$ , with a median of 0.04. The median OLR based on the surface area for COD equals 16.13 which is slightly less than the typical VF wetlands value of  $20 \text{ g COD m}^{-2}\cdot\text{d}^{-1}$ .

#### Substep a: Impact of quality categories

Figure 7 shows a plot of the removal performance vs. type of climate, according to the usual removal performance (Supplementary Table S2). Climates, as differentiated by and correlated with precipitation and air temperature (Supplementary



**Figure 7** | Removal performances in primary and HF as secondary treatment and VF wetlands combined, as a function of the type of climate, with typical performance ranges (horizontal dashed blue lines), from [Cross et al. \(2021\)](#). For BOD, the unique dashed line corresponds to mean performance, while for TSS, the unique dashed line corresponds to the minimum expected performance. The climates are extracted from [Chen & Chen \(2013\)](#): Am – Tropical monsoons, Aw – Tropical savannah with dry summers, BW – Desert (arid), BS – Steppe (semi-arid), Cf – Mild, temperate, fully humid, Cs – Mild, temperate with dry summers, Cw – Mild, temperate with dry winters, Df – Snow, fully humid, Dw – Snow with dry winters. The third letters denote: h – hot arid, a – hot summer, b – warm summer and c – cool summer.

Figure S10), may directly influence the nature-based processes occurring in TW. We have also observed lower or extreme removal performance in HF wetlands associated with ‘unsuitable’ labels for the following climates:

- Aw: Tropical savanna with dry winter, for cases in the USA (25 °C).
- Cfb: Mild temperate, fully humid, for cases in Kenya (19 °C) and Italy (14 °C).
- Csa, Csb: Mild temperate with dry summer, for cases in Jordan (15 °C), Chile (13 °C), Portugal (14 °C) and Spain (18 °C).
- Dfb: Snow, fully humid, for cases in Poland (8 °C).

We observed lesser performance for the ‘Other mineral’ substrate category, comprising fine and coarse volcanic tuff, tezontle extrusive rock (corresponding to Mexico, Csa, 23 °C and Jordan) and steel slag (Brazil, Cfa, 20 °C) for COD, BOD and TSS removals (Supplementary Figure S13). We cannot exclude the effect of climate.

### Substep b: Identification of correlations and causality

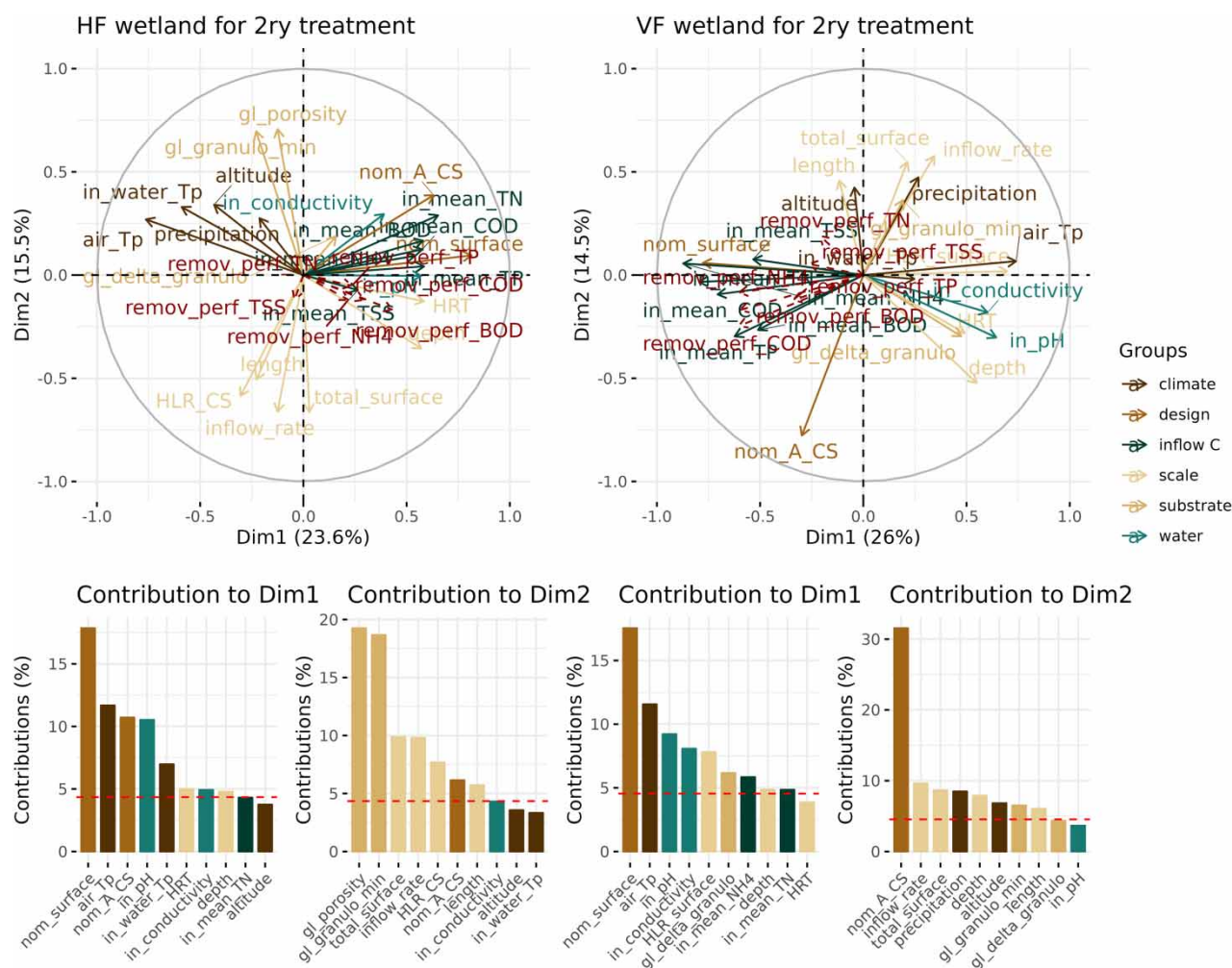
The following discussion will only present results for the second imputed dataset from application of the MICE method (Supplementary Figures S6 and S7). For HF wetlands, our imputation using the modified predictor matrix displayed fewer



changes in the initial correlation matrix (presence of correlations and value), compared to the standard predictor matrix. The removal of unsuitable values did lead to a reduction in some of the weaker correlations (Figure S8). For VF wetlands, the modified matrix also led to fewer changes, namely weakened strong correlations and amplified weak correlations. The removal of unsuitable values did weaken strong correlations for the initial non-imputed dataset.

**Principal components (PCs).** Using MFA, 36 and 40% of the variance in the datasets can be explained by the first two principal components (Figure 8). Variables within a common group are on the whole correlated with one another. Removal performance tends to follow inflow concentrations. For HF wetlands, the variables that mainly contributed to PC1 are related to climate, design and influent concentrations, while those that mainly contributed to PC2 are related to scale, substrate and water characteristics (pH and conductivity). PC3 comprises 12% of variance explained by substrate, water and scale, and PC4 comprises 11%, explained mainly by water.

For VF wetlands, the climate, design and influent concentrations mainly contribute to PC1, while design and scale mainly contribute to PC2. Interestingly, the corresponding quantitative variables used for substrate and scale did not discriminate among the various categories (Supplementary Figure S10). PC3 comprises 11% of the variance explained by substrate and climate, and PC4 comprises 10%, explained mainly by the substrate.



**Figure 8** | Distribution of variance of the imputed HF and VF wetlands with primary treated wastewater, yet without considering removal performance, along two principal components and with variables being grouped into broader categories. The longer the arrows, the larger the variable contribution is. Removal performance (in red) was used as a supplementary variable and did not therefore participate in the components explaining the variance.

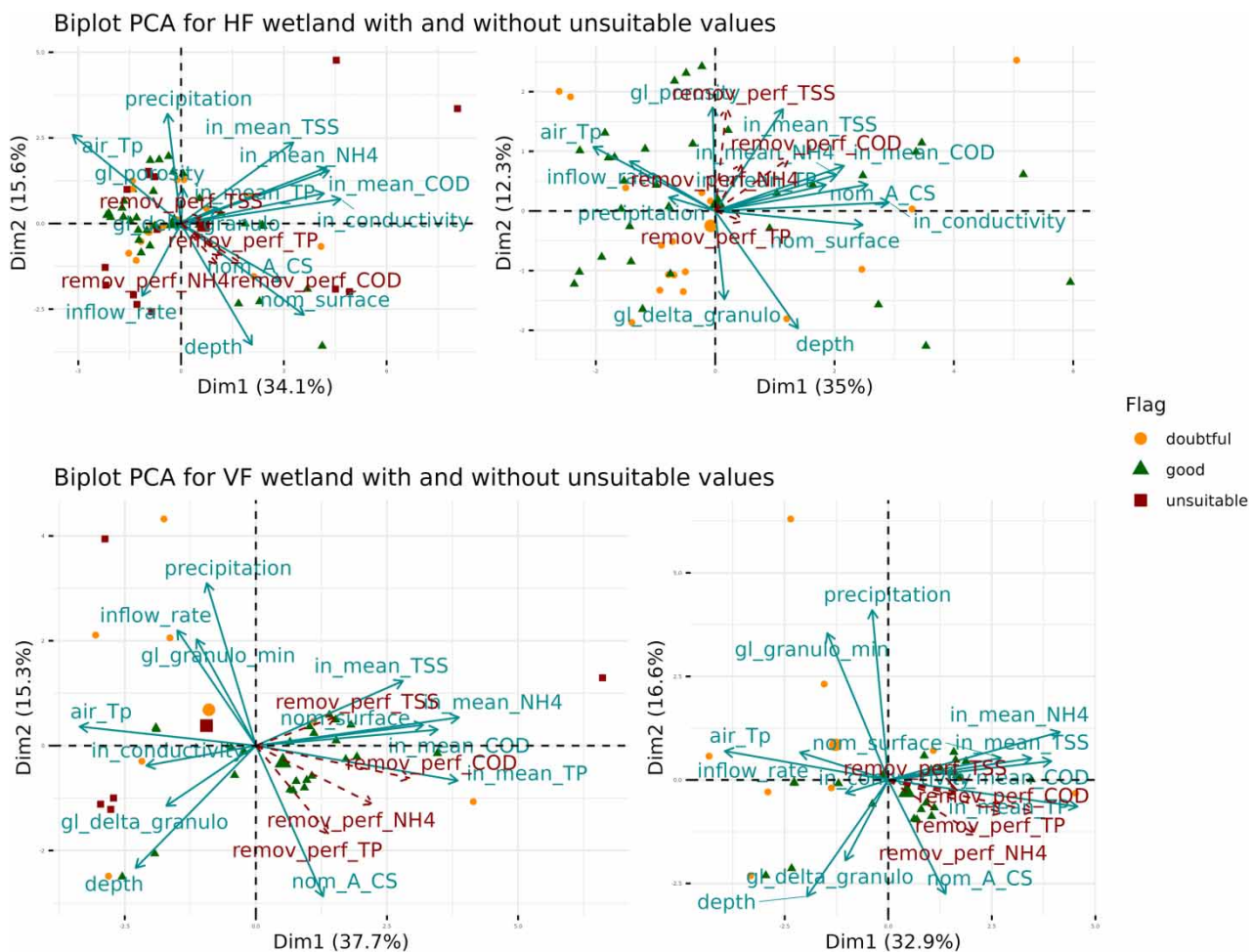


For PCA, we discarded from the principal component computation those variables showing a correlation of above 0.5:

- The influent water temperature was removed, air temperature was retained, influent pH was removed and influent conductivity was retained.
- The general porosity was removed and general minimum granulometry retained.
- HLR, length and total surface area were removed, the inflow rate retained, HRT removed and 'nominal' surface area (=total surface area/inflow rate) retained.
- The BOD and TSS influent concentrations were removed, and COD, NH<sub>4</sub>-N, TN and TP were all retained (see correlation plots in Supplementary Figure S8).

The first two components can thus explain 45% of the variance for HF wetlands and 50% for VF wetlands (Figure 9). Influent concentrations are the main contributors to PC1, while variables related to scale contribute to PC2. The detailed contributions of each variable to the first four PCs are available in Supplementary Figure S9. From a visual perspective, it can be observed that the removal of 'unsuitable' sampling campaigns has allowed for the removal of the extreme observations that had been driving some of the correlations. The variance distribution is similar overall to the MFA results.

The removal performances (red) are on the whole correlated with one another as well as with the influent concentrations. In the case of HF wetland, the *Tank-in-series-k - C\** model, a simplified representation of hydraulics and pollutant kinetics, already includes this correlation, with the presence of a background residual concentration  $C^*$  such that it increases with the



**Figure 9** | PCA representing the variables and observations on the imputed HF as secondary treatment and VF wetlands, both before and after the removal of 'unsuitable' sampling campaigns. The longer the arrows, the larger the variable contribution is. We used the removal performance (in red) as a supplementary variable; thus, it does not participate in the components explaining the variance.

influent concentration, while the proportion  $C^*/C_{in}$  decreases as the influent concentration increases (Kadlec & Wallace 2009). The short length of the red arrows for  $NH_4$  and TP in Figure 9 indicates low correlations with the principal components and corresponding variables. For HF wetlands, the removal performances are aligned along PC1, to which the design variables contribute most. Hence, the larger the ‘nominal’ surface area and cross-sectional area (i.e. the larger the area per unit influent volume), the higher the removal performance is. For VF wetlands, the depth and ‘nominal’ cross-sectional area contribute most to PC2. Overall, the visualisation and results from the MFA and PCA should be considered with caution since half of the variance is not explained by the independent variables.

Surprisingly, air temperature and precipitation are negatively correlated and uncorrelated with removal performance. Conversely, it is well known that air temperature is positively correlated with and enhances reaction kinetics, as notably taken into account in the kinetic removal rate  $k$  as a simplified expression of the Arrhenius Law in Kadlec & Wallace (2009) and Von Sperling *et al.* (2023a, 2023b).

**Confounding factor.** The ‘nominal’ surface area (divided by the inflow rate) is negatively correlated with air temperature (as well as with water temperature): from  $-0.51$  with to  $-0.46$  without any ‘unsuitable’ sampling campaigns for HF wetlands. This correlation with temperature is also visible via the HRT: from  $-0.42$  to  $-0.48$  (Supplementary Figure S8). We consider that the removal of unsuitable sampling campaigns only barely modifies this correlation. A higher temperature normally enhances removal kinetic rates and treatment performances, so the correlation with surface rather than treatment performances suggests the existence of a causal relationship: knowing that higher temperature enhances treatment performance, the TW design is adapted in consequence. Aiming for a ‘standard’ performance, it is possible to under-design TW in warmer climates, while it is necessary to over-design TW to compensate colder climates. Therefore, as temperature increases, surface area decreases. As an example, Poland case-study features a considerably larger surface area-to-inflow rate ratio than other case studies. So in this case, the temperature effect is already being mitigated by the surface, which could explain the absence of any clear trend in removal performance by the type of climate. Air temperature, therefore, is a confounding factor in the causal relationship between surface area and removal performance. Hence, the impact of air temperature may have been mitigated (concealed) by this design adaptation.

The assumption that the compiled set of TW is indeed representative may prove to be optimistic since many factors actually influence TW functioning. For modelling purposes, using all variables present would lead to an inadequate SFR (or number of degrees of freedom). Further work is needed to determine all features relevant to TW removal performance.

To ensure reliability, we had only considered peer-reviewed publications. National databases and operational reports could also be included in order to increase the size of our dataset. The quality framework from Gootzen *et al.* (2023) provides valuable insights to both assess the quality of data sources and combine datasets for a given context (as defined by a target variable, target population and aggregation level). This framework takes into consideration the following dimensions, which would prove to be useful in our Steps 1 and 2: Relevance, Population coverage, Population representativeness, Variable validity, Concept stability, Correctability, Recentness, Process timing, Accessibility, Meta-data and Comparability.

## CONCLUSION

This work has collected and analysed design and treatment performance data on HF and VF wetlands for wastewater, with a literature-based approach and a worldwide scope. The size of the dataset obtained for this study demonstrated that a greater set of information and data should be reported in published results to facilitate interpretation and reuse.

In the perspective of data-driven modelling, we developed a general six-step methodology for data quality validation and moreover illustrated the process on a dataset on HF and VF wetlands. Assuming continuity between ‘outliers’, ‘faulty’ and ‘good’ data is a cornerstone of the reasoning process, allowing us to better comprehend the complexity of data validation. The quality assessment output is less categorical than the typical ‘good’, ‘doubtful’ and ‘bad’ labels. The quality output can be considered and mobilised directly as an importance weight for each observation to improve the reliability of data-driven models. In accordance with this approach, we analysed some customary data handling practices and provided suggestions of adequate indicators, metrics and outputs for greater clarity, comparability and auditability.

The ‘automation’ of data validation and quality labelling should not be understood as a way to dispossess experts from decision making. On the opposite, the inputs of the experts are needed in the suggested iterative data validation workflow. Expert knowledge is required to identify variables and key factors, define categories of equivalence, select reference knowledge and thresholds, and the construct checks and quality categories.

Finally, it should be stressed that the diversity of case studies limits the power of the quality checks, in particular the expert knowledge chosen for checking a global population of TW may not be as relevant as finer and more localised checks. A better characterisation of subpopulations of TW can participate in defining more reliable quality checks and models.

Despite over-simplified (e.g. rule of thumb) guidelines on TW conception, we could observe that temperature was a key factor already considered in TW designs to achieve adequate pollution treatment. The correlation-based MFA and PCA visually showed that the removal performance of each pollutant depended on different variables, related not only to design but also to climate, system scale, initial wastewater composition and substrate.

Last, we feel that including this methodology in research practices will allow for a more efficient production and (re)use of data, in addition to leading to more reliable data-driven models, along with a common data exchange format.

## ACKNOWLEDGEMENTS

Special thanks are due to: Damien Tedoldi and Jean-Luc Bertrand Krajewski (INSA DEEP), who provided insights for the Data Validation methodology, Lluís Bosch, who implemented the database on the ICRA server, Janick Klink for explaining the handling of raster formats, Javier Ortiz-Rivero, who provided insights for the Exploratory Data Analysis and Lide Jaurrieta (ICRA), who made the illustration of the treatment wetland in the graphical abstract.

## FUNDING

This project is part of the MULTISOURCE European project, which has received funding from the European Union's Horizon H2020 innovation action programme under grant agreement 101003527. This particular research, as a joint venture between INRAE and ICRA, has also received funding from H2O'Lyon University's School of Research on Water and Hydro-systems Sciences as well as from the Economy and Knowledge Department of the Catalan Government through Consolidated Research Groups ICRA tech (2021-SGR-01283) and LEQUIA (2021-SGR-01352).

## DATA AVAILABILITY STATEMENT

All relevant data are available from an online repository or repositories: <https://forgemia.inra.fr/reversa/nature-based-solutions/multisource/tw.dataset>.

## CONFLICT OF INTEREST

The authors declare there is no conflict.

## REFERENCES

- Acuña, V., Castañares, L., Castellar, J., Comas, J., Cross, K., Istenic, D., Masi, F., McDonald, R., Pucher, B., Pueyo-Ros, B., Riu, A., Rizzo, A., Riva, M., Tondera, K. & Corominas, L. 2023 *Development of a decision-support system to select nature-based solutions for domestic wastewater treatment*. *Blue-Green Systems* **5** (2), 235–251. <https://doi.org/10.2166/bgs.2023.005>.
- Buuren, S. v., 2018 *Flexible Imputation of Missing Data*, 2nd edn. Chapman & Hall/CRC Press, Boca Raton, FL, USA. Available from: <https://stefvanbuuren.name/fimd/>.
- Buuren, S. V. & Groothuis-Oudshoorn, K. 2011 *Mice: Multivariate imputation by chained equations in R*. *Journal of Statistical Software* **45** (3). <https://doi.org/10.18637/jss.v045.i03>.
- Canguilhem, G. 1966 *Le Normal Et Le Pathologique*.
- Canler, J.-P. & Perret, J.-M. 2007 *FND AE N35: Les Clari-Floculateurs*. Available from: [http://www.fndae.fr/documentation/PDF/fndae35\\_a.pdf](http://www.fndae.fr/documentation/PDF/fndae35_a.pdf).
- Chan, F. K. S., Griffiths, J. A., Higgitt, D., Xu, S., Zhu, F., Tang, Y.-T., Xu, Y. & Thorne, C. R. 2018 'Sponge city' in China – A breakthrough of planning and flood risk management in the urban context. *Land Use Policy* **76** (July), 772–778. <https://doi.org/10.1016/j.landusepol.2018.03.005>.
- Chen, D. & Chen, H. W. 2013 *Using the Köppen classification to quantify climate variation and change: An example for 1901–2010*. *Environmental Development* **6** (April), 69–79. <https://doi.org/10.1016/j.envdev.2013.03.007>.
- Choudhary, A. K., Kumar, S. & Sharma, C. 2011 *Constructed Wetlands: An Approach for Wastewater Treatment*, 8.
- Clemens-Meyer, F., Bertrand-Krajewski, J. L. & Lepot, M. 2021 *Metrology in Urban Drainage and Stormwater Management: Plug and Pray*.
- Corominas, L., Garrido-Baserba, M., Villegz, K., Olsson, G., Cortés, U. & Poch, M. 2018 *Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques*. *Environmental Modelling & Software* **106** (August), 89–103. <https://doi.org/10.1016/j.envsoft.2017.11.023>.

- Cross, K., Tondera, K., Rizzo, A., Andrews, L., Pucher, B., Istenic, D., Karres, N. & McDonald, R. 2021 *Nature-Based Solutions for Wastewater Treatment – A Series of Factsheets and Case Studies*. IWA Publishing, London, UK.
- Cunha, A. S. d., Peixoto, F. C. & Prata, D. M. 2021 *Robust data reconciliation in chemical reactors*. *Computers & Chemical Engineering* **145** (February), 107170. <https://doi.org/10.1016/j.compchemeng.2020.107170>.
- Desrosières, A. 1992 *Séries Longues Et Conventions d'équivalence*. *Genèses* **9** (1), 92–97. <https://doi.org/10.3406/genes.1992.1665>.
- Desrosières, A. 2006 *From Courtot to Public Policy Evaluation: Paradoxes and Controversies Involving Quantification*.
- Feinerer, I., Hornik, K. & Meyer, D. 2008 *Text mining infrastructure in R*. *Journal of Statistical Software* **25** (5). <https://doi.org/10.18637/jss.v025.i05>.
- Fonder, N. & Headley, T. 2013 *The taxonomy of treatment wetlands: A proposed classification and nomenclature system*. *Ecological Engineering* **51** (February), 203–211. <https://doi.org/10.1016/j.ecoleng.2012.12.011>.
- Gootzen, Y. A. P. M., Daas, P. J. H. & Van Delden, A. 2023 *Quality framework for combining survey, administrative and big data for official statistics*. *Statistical Journal of the IAOS* **39** (2), 439–446. <https://doi.org/10.3233/SJI-220110>.
- Hauduc, H. 2011 *Modèles Biocinétiques de Boues Activées de Type ASM: Analyse Théorique Et Fonctionnelle, Vers Un Jeu de Paramètres Par Défaut*. PhD Thesis.
- Henze, M., van Loosdrecht, M. C. M., Ekama, G. A. & Brdjanovic, D. 2008 *Biological Wastewater Treatment: Principles, Modelling and Design*. IWA Publishing, London, UK. <https://doi.org/10.2166/9781780401867>.
- Jones, E. R., Van Vliet, M. T. H., Qadir, M. & Bierkens, M. F. P. 2021 *Country-level and gridded estimates of wastewater production, collection, treatment and reuse*. *Earth System Science Data* **13** (2), 237–254. <https://doi.org/10.5194/essd-13-237-2021>.
- Kadlec and Wallace 2009 *Treatment Wetlands*.
- Karger, D. N., Conrad, O., Böhrer, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Peter Linder, H. & Kessler, M. 2017 *Climatologies at high resolution for the earth's land surface areas*. *Scientific Data* **4** (1), 170122. <https://doi.org/10.1038/sdata.2017.122>.
- Kassambara, A. & Mundt, F. 2020 *Factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. Available from: <https://CRAN.R-project.org/package=factoextra>.
- Langergraber, G. 2011 *Numerical modelling: A tool for better constructed wetland design?* *Water Science and Technology* **64** (1), 14–21. <https://doi.org/10.2166/wst.2011.520>.
- Langergraber, G., Dotro, G., Nivala, J., Rizzo, A. & Stein, O. R. 2020 *Wetland Technology: Practical Information on the Design and Application of Treatment Wetlands*. IWA Publishing. <https://doi.org/10.2166/9781789060171>.
- Langergraber, G., Castellar, J. A. C., Andersen, T. R., Andreucci, M.-B., Baganz, G. F. M., Buttiglieri, G., Canet-Martí, A., Carvalho, P. N., Finger, D. C., Bulc, T. G., Junge, R., Megyesi, B., Milosevic, D., Oral, H. V., Pearlmutter, D., Pineda-Martos, R., Pucher, B., van Hullebusch, E. D. & Atanasova, N. 2021a *Towards a cross-sectoral view of nature-based solutions for enabling circular cities*. *Water* **13** (17), 2352. <https://doi.org/10.3390/w13172352>.
- Langergraber, G., Castellar, J. A. C., Pucher, B., Baganz, G. F. M., Milosevic, D., Andreucci, M.-B., Kearney, K., Pineda-Martos, R. & Atanasova, N. 2021b *A framework for addressing circularity challenges in cities with nature-based solutions*. *Water* **13** (17), 2355. <https://doi.org/10.3390/w13172355>.
- Lê, S., Josse, J. & Husson, F. 2008 *Factominer: A package for multivariate analysis*. *Journal of Statistical Software* **25** (1), 1–18.
- Masi, F., Rizzo, A. & Regelsberger, M. 2018 *The role of constructed wetlands in a new circular economy, resource oriented, and ecosystem services paradigm*. *Journal of Environmental Management* **216** (June), 275–284. <https://doi.org/10.1016/j.jenvman.2017.11.086>.
- Mercoiret, L. 2009 *Qualité Des Eaux Usées Domestiques Produites Par Les Petites Collectivités – Application Aux Agglomérations d'assainissement Inférieures à 2 000 Equivalent Habitants*, 55.
- Muhr, D., Affenzeller, M. & Küng, J. 2023 *A probabilistic transformation of distance-based outliers*. *Machine Learning and Knowledge Extraction* **5** (3), 782–802. <https://doi.org/10.3390/make5030042>.
- Oliveira, S. C., Souki, I. & Von Sperling, M. 2012 *Lognormal behaviour of untreated and treated wastewater constituents*. *Water Science and Technology* **65** (4), 596–603. <https://doi.org/10.2166/wst.2012.899>.
- Ooms, J. 2023 *Pdftools: Text Extraction, Rendering and Converting of PDF Documents*. Available from: <https://docs.ropensci.org/pdftools/> (website) <https://github.com/ropensci/pdftools#readme> (devel) <https://poppler.freedesktop.org> (upstream).
- Oral, H. V., Carvalho, P., Gajewska, M., Ursino, N., Masi, F., van Hullebusch, E. D., Kazak, J. K., Exposito, A., Cipolletta, G., Andersen, T. R., Finger, D. C., Simperler, L., Regelsberger, M., Rous, V., Radinja, M., Buttiglieri, G., Krzeminski, P., Rizzo, A., Dehghanian, K., Nikolova, M. & Zimmerman, M. 2020 *A review of nature-based solutions for urban water management in European circular cities: A critical assessment based on case studies and literature*. *Blue-Green Systems* **2** (1), 112–136. <https://doi.org/10.2166/bgs.2020.932>.
- Pebesma, E., Mailund, T. & Hiebert, J. 2016 *Measurement units in {R}*. *R Journal* **8** (2), 486–494. <https://doi.org/10.32614/RJ-2016-061>.
- Pucher, B. & Langergraber, G. 2019 *The state of the art of clogging in vertical flow wetlands*. *Water* **11** (11), 2400. <https://doi.org/10.3390/w11112400>.
- Ramasamy, J., Devanathan, S. & Jayaraman, D. 2021 *Comparative analysis of select techniques and metrics for data reconciliation in smart energy distribution network*. *Water Supply* **21** (5), 2109–2121. <https://doi.org/10.2166/ws.2020.314>.
- Rieger, L. 2012 *Guidelines for using activated sludge models*. *Water Intelligence Online* **11** (September). <https://doi.org/10.2166/9781780401164>.
- Rieger, L., Takács, I., Villeg, K., Siegrist, H., Lessard, P., Vanrolleghem, P. A. & Comeau, Y. 2010 *Data reconciliation for wastewater treatment plant simulation studies-planning for high-quality data and typical sources of errors*. *Water Environment Research* **82** (5), 426–433. <https://doi.org/10.2175/106143009X12529484815511>.



- Rousseau, D. P. L., Vanrolleghem, P. A. & De Pauw, N. 2004 *Model-based design of horizontal subsurface flow constructed treatment wetlands: A review*. *Water Research* **38** (6), 1484–1493. <https://doi.org/10.1016/j.watres.2003.12.013>.
- Team, R Core 2023 *R: A Language and Environment for Statistical Computing*. Available from: <https://www.R-project.org/>.
- United Nations Environment Programme 2023 *Nature-Based Infrastructure: How Natural Infrastructure Solutions Can Address Sustainable Development Challenges and the Triple Planetary Crisis*. United Nations Environment Programme. <https://doi.org/10.59117/20.500.11822/44022>.
- Villez, K. 2017 Data reconciliation with inequality constraints induces bias: A cause for concern? In *AIChE Annual Meeting*, p. 6.
- Von Sperling, M., Verbyla, M. E. & Oliveira, S. M. A. C. 2020 *Assessment of Treatment Plant Performance and Water Quality Data: A Guide for Students, Researchers and Practitioners*. IWA Publishing, London, UK. <https://doi.org/10.2166/9781780409320>.
- Von Sperling, M., Wallace, S. D. & Nivala, J. 2023a *What is the best procedure for determining removal rate coefficients in horizontal flow treatment wetlands: Influent and effluent concentrations or longitudinal concentration profiles?* *Water Science & Technology* **87** (10), 2541–2552. <https://doi.org/10.2166/wst.2023.144>.
- Von Sperling, M., Wallace, S. D. & Nivala, J. 2023b *Representing performance of horizontal flow treatment wetlands: The tanks in series (TIS) and the plug flow with dispersion (PFD) approaches and their application to design*. *Science of The Total Environment* **859** (February), 160259. <https://doi.org/10.1016/j.scitotenv.2022.160259>.
- Wallace, S. & Knight, R. 2005 *Small-Scale Constructed Wetland Treatment Systems Feasibility, Design Criteria, and O&M Requirements*. WERF. Available from: <https://partage.inrae.fr/service/home/~/?auth=co&loc=fr&id=16600&part=2>.
- Zhu, J.-J., Yang, M. & Ren, Z. J. 2023 *Machine learning in environmental research: Common pitfalls and best practices*. *Environmental Science & Technology*. June, acs.est.3c00026. <https://doi.org/10.1021/acs.est.3c00026>.

First received 22 March 2024; accepted in revised form 20 May 2024. Available online 5 June 2024