Water Practice & Technology

IWA PUBLISHING

Water Practice & Technology Vol 17 No 10, 1977 doi: 10.2166/wpt.2022.113

One script to solve it all – an open-source-based framework for a digital workflow based on WWTP data

Markus Ahnert [a,*] and Stefan Hurzlmeier[b]

[a] Institute of Urban and Industrial Water Management, Technische Universität Dresden, 01062 Dresden, Germany
[b] ZWT Engineering GmbH, Gottlieb-Keim-Straße 28, 95448 Bayreuth, Germany
*Corresponding author. E-mail: markus.ahnert@tu-dresden.de

MA, 0000-0001-9033-1847

ABSTRACT

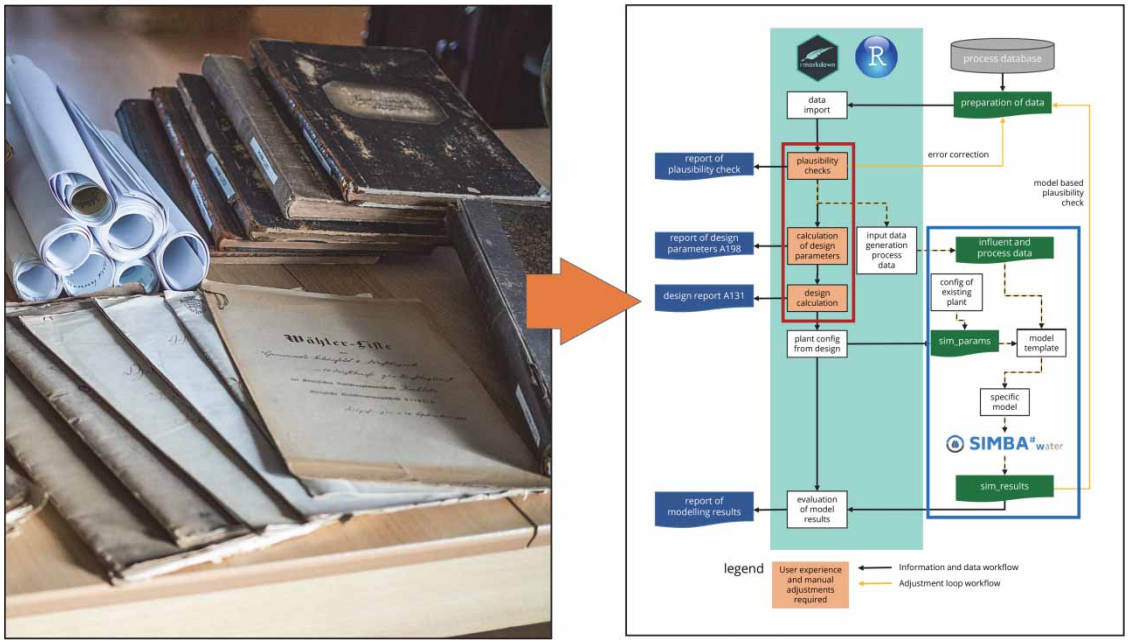
The increasing use of digital technology has many advantages, but is insufficiently used in urban water management. This article describes a workflow based on detailed German rules and regulations for the use of routine data from wastewater treatment plants (WWTP) for design, assessment, benchmarking, and activated sludge modelling. Our workflows use data that are routinely collected. However, prior to use the data must first be checked. It includes data import, plausibility checks, data evaluation, and subsequent plant design. In addition, the data available in the system can be used to supply an activated sludge model of the plant as a basis for use in redesign and process optimisation. The main advantages of such an approach are time savings, increased quality, and a transparent and comprehensible procedure. Exemplary results are presented for individual elements of the workflow. We also provide a comprehensive classification of the question in the context of environmental data science. WWTP routine data bases are an important source for information needed for evaluation and optimisation.

Key words: design, digital workflow, environmental data science, open source, routine data

HIGHLIGHTS

- Classification of WWTP routine data in the context of environmental data science.
- First description of a complete digital workflow for assessment of routine data.
- Modular system for future extension.
- Open-source based framework with possible combination with other software tools (open-source or commercial).
- Joint use of calculation and documentation.

GRAPHICAL ABSTRACT



INTRODUCTION

In many areas, data plays a crucial role in the planning, dimensioning and operational optimisation of plants. This is particularly true in the environmental field, because only by knowing the loads and the resulting treatment performance in treatment plants can the environmental objective be ensured. This article deals with the handling of data from municipal wastewater treatment plants. Depending on the type of data, the acquisition of operational data is either automated by monitoring processes in the SCADA system (supervisory control and data acquisition) or by manual entry, e.g. of laboratory values. Particularly in Germany, but found in comparable form worldwide, is the documentation of the routine data collected in operating journals. In addition to being stored in intervals of seconds or minutes, the existing data is aggregated, e.g. as daily mean values. These data were used for documentation, evaluation, and new design of the plants. The quality of the data has a major impact on all uses based on it. Systematic and transparent evaluation can ensure the usefulness of the data for subsequent uses. The DWA regulations (German Association for Water, Wastewater and Waste) relevant for the treatment plant design in Germany, Austria, and Switzerland are based on routine data available in the form described.

FUNDAMENTALS AND REQUIREMENTS

This type of data collection, preparation and further processing is a branch of environmental data science (EDS). This term is relatively new and is not found much in the literature (Scopus® 17 entries; Web of Science® 18 entries, as of December 2021). Mostly, the term is used in the field of 'big data'; the much better-known term that is now used in countless publications.

In addition to a definition of EDS, Gibert *et al.* (2018) provide a series of challenges that thematically fit the question examined in this article. The numbering of the challenges is based on the article by Gibert *et al.* (2018).

Challenge 2: Lack of data science skills within the environmental science curricula

As can also be seen from the two illustrations in Figure 1 and supported by the statements of Gibert *et al.* (2018), data science is missing from many topics of environmental science. In some environmental research groups, the need to include other disciplines has been recognised, so that experts in computer science, mathematics and statistics, for example, are integrated into the research in order to meet the requirements of the complexity and scope of the data.

Classical settings, i.e. in smaller planning offices or operators of wastewater associations, usually contain experts. They tend to have mathematical and statistical knowledge (MSK) – also due to the proximity to traditional training concepts. They apply this in their daily work. With the appropriate professional background

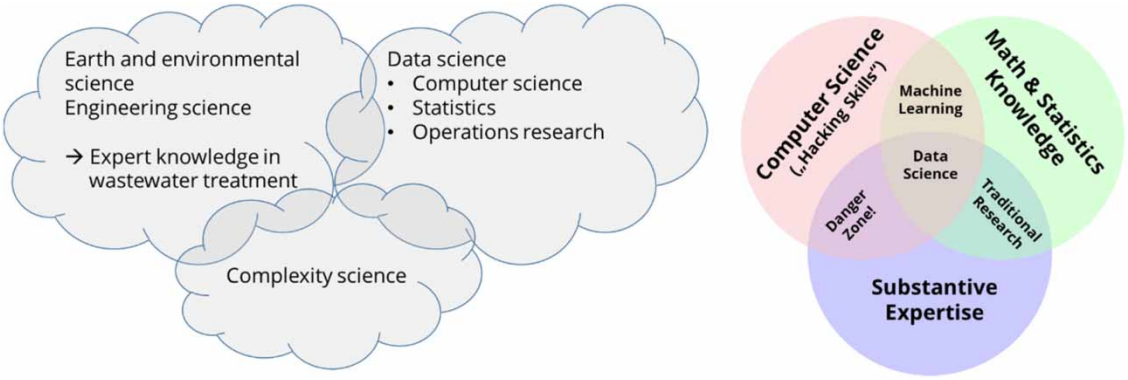


Figure 1 | Left: The cross-disciplinary nature of environmental data science in the field of wastewater treatment (adapted from Blair *et al.* 2019). Right: Data science venn diagram (modified from Conway 2013).

or personal inclination, further development in the direction of data science is possible. Furthermore, collaboration with experts from the fields of 'Computer Science' or 'MSK' is also conceivable, so that more complex tools and methods can also be used. However, from the authors' experience, this seems to be the exception rather than the rule.

Challenge 3: Methodological gaps for designing data science processes in real applications

In a literature survey, no guidelines for the design of workflows suitable for our problem could be found. This article is also intended to be a first step in this direction and to point out special features in the practical application.

Challenge 5: Data quality and dealing with uncertainty in data – in combination with Challenge 7: Methods to choose pertinent, correct, sufficient, and non-superfluous data for analysis

These challenges are currently most recognised in the field of EDS. A broad base of methods and tools are available. However, there are still deficits in the selection and testing of applicability and actual use in daily engineering practice.

Challenge 11: Reproducibility and interoperability

This challenge is the main concern of this article. It has emerged from our own research and engineering practice that there are currently major deficits in designing data aggregation and evaluation in a transparent way so that reproducibility is ensured. The advantages of so doing are typically not perceived as such, but only possible disadvantages and negative effects such as the passing on of knowledge or the competitive disadvantage due to the supposed transfer of knowledge to competitors.

Blair *et al.* (2019) define NDS in a more concrete way than many other authors, who mostly use the term 'big data'. They see the complexity and heterogeneity of the data and data sources used as the main challenge in dealing with environmental data, not the pure mass of data.

These authors structure the challenges they identify somewhat differently than Gibert *et al.* (2018), but the basic statements are very similar. Below is a comparison of the challenges of Blair *et al.* (2019) in the context of the conditions of this article:

- **Data challenge:** As explained above, it is not necessarily the mass of data that poses the greatest challenge. Therefore, even 'smaller' data volumes, such as in the present case of the operating journal of wastewater treatment plants based on daily data records, can make evaluation and use difficult (for information: 365 data sets per year and, depending on the complexity of the plant, up to 100 data points). In Germany, sampling from self-monitoring usually only takes place on a few days of the week, so that there are large data gaps in some measured values. Due to weekly variation effects, this can have a considerable influence on evaluations of these data.
- **Modelling challenge:** In the field of wastewater treatment, process models are mostly used (e.g. activated sludge models Henze *et al.* (2000)). These require data as model input to describe the influent load and other conditions (e.g. temperature) as well as comparative data in the mass flows leaving the plant (effluent, sludge

discharge) and internal comparative variables (e.g. solids concentrations in the reactors). These data must be checked for their suitability, but a direct development of data-driven models is not necessary for this.
- **The spatial/temporal challenge:** The spatial resolution of wastewater treatment plants is not a problem due to the defined measuring points. The highly heterogeneous temporal resolution (e.g. samples of chemical analyses at intervals of several days compared to high-resolution signals from online sensors with an interval of minutes or seconds = continuous data; US-EPA 2015) of the data used is a special problem (see also Newhart *et al.* 2019).
- **Complexity challenge:** Due to the diversity of the processes involved in wastewater treatment and the internal mass flows, wastewater treatment plants are complex systems. In addition to the challenges in the analysis of such plants and processes, there are also advantages, e.g. in the evaluation of the effects of process adjustments on the overall plant behaviour.
- **Uncertainty challenge:** All data are subject to a wide variety of influencing factors during measurement and further processing. The resulting uncertainties can massively influence the conclusions derived from the data, so that a detailed examination and evaluation is necessary in the data processing process.
- **Cross-disciplinary challenge:** Handling the required data requires suitable experts, tools, and methods.

Experience shows that the systematic consideration of the principles and the application of methods of EDS in the field of wastewater treatment are not yet used. A literature search revealed only a few approaches, as presented in the Material and Methods section. Since the tools and methods developed for big data have their strengths in the analysis of large data volumes that are available in high temporal resolution, their application is more useful in the field of process optimisation (see e.g. Qiu *et al.* 2018). The use for WWTP design plays only a minor role.

Based on the challenges described above, a framework for standardised data evaluation and further processing for the design of WWTP was developed, which addresses the following goals:

- Use of EDS for semi-automated design of WWTPs
- Continuous workflow for data aggregation, evaluation, plausibility check, design, and process optimisation (including activated sludge modelling)
- Transparent digital documentation
- Use of open source tools for data workflow and for linking commercial software packages
- Comprehensible, repeatable, reusable or easily adaptable methodology

The implementation is based on a case study using the DWA rules and regulations for the design of municipal WWTP. In the design rules, the development of the design basis and the process-related design of wastewater treatment plants is specified in great detail. The rules and regulations are suitable for deriving algorithms for automated calculations. However, this was not the aim of the development of the rules; consequently some aspects in the instructions remain unclear. This is discussed during the implementation in a digital workflow.

The description of the procedure, the hurdles to be overcome and the initial experiences from the application are given in the following sections.

MATERIAL AND METHODS

Background to the case study

The DWA rules and regulations represent the generally recognised codes of practice in Germany, Austria, and Switzerland. Other countries have also adopted these guidelines or adapted parts of it. Some important regulations are available in English (e.g. Worksheet A 131 (DWA-A131 2016) in a version adapted for other climate zones (DWA 2018).

In this set of rules, there are very specific calculations for the process engineering design of systems. The basis for this is appropriately prepared data that is routinely collected as part of self-monitoring. The sampling frequency within the scope of self-monitoring is determined in Germany by federal regulations and depends on the size group of the wastewater treatment plant.

The determination and processing of this data is also partly specified in the regulations. The greatest influence is exerted by the combination of regulations for determining the design data for wastewater treatment plants A 198 (ATV-DVWK-A198E 2003) and for the design of activated sludge plants A 131 (DWA-A131 2016). These regulations are applied for the design of approximately 9100 wastewater treatment plants (Federal Statistical

Office of Germany; as of 2016) in Germany alone. Figure 2 shows the distribution of these wastewater treatment plants by size class. While about one third represent rather simple plants with up to 1000 PE (people equivalent; 1 PE = 60 g BOD_5 $person^{-1}$ d^{-1}), a more complex process management is to be expected for about one quarter of the plants, where the dynamic simulation of the treatment processes can also be useful in the context of optimisation tasks. Therefore, it seemed reasonable to develop a continuous workflow for this process, which can be expanded modularly according to the tasks.

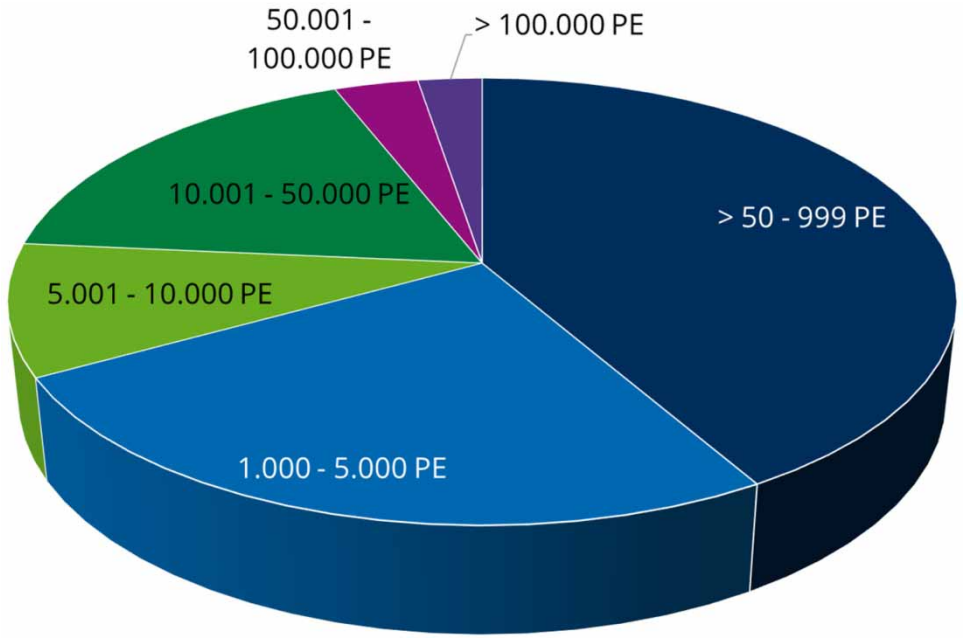


Figure 2 | Distribution of wastewater treatment plants in Germany by size (based on influent load).

This essentially consists of various modules to be processed one after the other, some of which are arranged in iterative loops. The calculations themselves are simple time series analyses combined with the formation of mean or percentile values over different time periods. These calculations can be carried out with various calculation tools, e.g. spreadsheets, statistical software packages, mathematical calculation software, or dedicated software solutions (e.g. DATAR by Gabaldón *et al.* (1998) or DESASS by Ferrer *et al.* (2008)). In their review of WWTP design for the pulp and paper industry, Sonaje & Berlekar (2015) show a wide range of software solutions more focused on modelling.

A first essential step is to ensure data quality. Exploratory data analysis methods lend themselves to this, but they tend to be of secondary importance in EDS (see Corominas *et al.* 2018; Newhart *et al.* 2019). The latter describe this very impressively:

> '*Simple descriptive statistics also belong to this level, as this is how data is managed today in WWTPs. However, we did not find in literature scientific papers dealing exclusively with simple descriptive statistics.*' (Corominas et al. 2018; page 4)

An integrated workflow has many advantages, as described by Blischak *et al.* (2019). The authors present an open source framework for reproducible research. While Blischak *et al.* (2019) focuses more on the technical aspects (programming language, code exchange platform, automated documentation), Lowndes *et al.* (2017) explain the temporal stages on the way to a more sustainable and reproducible way of conducting research activities and publications. The resulting benefits are presented in detail.

In the field of building information modelling (BIM; ISO_29481-1 2016), a few ideas and concepts can be found that are related to the wastewater treatment sector. Marzouk & Othman (2017) describe a basic approach using an example of anaerobic ponds, without specifically addressing the design of the plant components. The implementation of BIM using Industry Foundation Classes (IFC ISO_16739-1 2018) is also illustrated with various examples (secondary clarifier (Söbke *et al.* 2018; Söbke *et al.* 2021) and anaerobic digestion reactors for

biogas plants (Söbke *et al.* 2020). The need for an extension of the IFC standard is intensively discussed. This is currently the only open standard for use in open BIM environments. Di Biccari & Heigener (2018) developed several extensions (reliable knowledge source, legal and technical regulations) and a semantic model to improve the usability of the IFC standard in the WWTP sector.

A more systemic approach is presented by Manig *et al.* (2019) for the future of design of WWTP. They do not give any concrete advice on how to carry out the calculations, instead they recommend a repeated use of procedures and methods that have to be re-selected and adapted in an iterative process due to the time-varying influencing factors.

The research reviewed above did not yield any solutions applicable to the need for a digital workflow in the field of WWTP design. Therefore, the sources presented were taken as a suggestion to develop a method for the case study of the design of wastewater treatment plants described above.

Development of the concept

The first step was to develop the requirements for the workflow:

- Modular structure (enables partial interchangeability), but requires
- Clearly defined interfaces combined with a predefined system, e.g. for the variables.
- Concrete description of the requirements for the input data
- Creation of output information suitable for the underlying calculation rules (sets of rules)
- Iteration loops for data cleaning or plausibility checks, documentation of the necessary changes to the data
- Final automated creation of a documentation of the data used and the calculated results.

In order to guarantee objectivity and transparency, the use of open source solutions is desirable. This entails disadvantages and reservations. Acceptance problems are conceivable because of fears that too much know-how will be revealed by consultants in the course of a design process. This is discussed later.

Technical implementation

Before starting, research was carried out into suitable programming languages. The languages python and R are very well suited and widely used. Since one of the main goals is an integrated automated documentation, advantages were seen in using R (R Core Team 2021) in combination with the libraries knitr (Xie 2018) and markdown (Allaire *et al.* 2021). This combination enables the automatic generation of documents in html, Word, or PDF format. Other similarly designed programming environments can also be used.

The interaction of the various tools and their use in the context of comprehensible research and application is described in detail by Stodden *et al.* (2014).

Based on the requirements profile described above and by extracting the necessary calculation specifications from the regulations, a flow chart was developed based on the modules developed and the software used with the libraries described (see Figure 3). The middle coloured area represents modules developed in R. The blue reports are documentation and the green blocks represent data sources in a spreadsheet format.

The data used are fed into the evaluation scripts via a spreadsheet interface. The area bordered in red represents the modules for the specific calculation in direct connection with the rules documents. The resulting findings and experiences are described and discussed later. All other modules arranged around it are necessary to prepare the data and to meet the quality requirements as well as for further use of the data in subsequent steps in the design process. An essential step is the creation and use of an activated sludge model in the software Simba# (ifak 2021). External programmes are controlled directly from the R script via API interfaces. The blue frame contains the necessary data and modules.

The engineering workflow is represented by the black arrows. The yellow arrows represent adjustment loops that are based on findings from the testing and design process and lead to adjustments to the data and information.

RESULTS

The workflow shown in Figure 3 was implemented in R and Rmarkdown. Data records from operating journals of wastewater treatment plants of various sizes were available as input data. These have different data density, for example the plant inlet has a size-dependent minimum sampling frequency according to legal requirements in Germany. The data covered periods of several years each in the form of daily mean values.

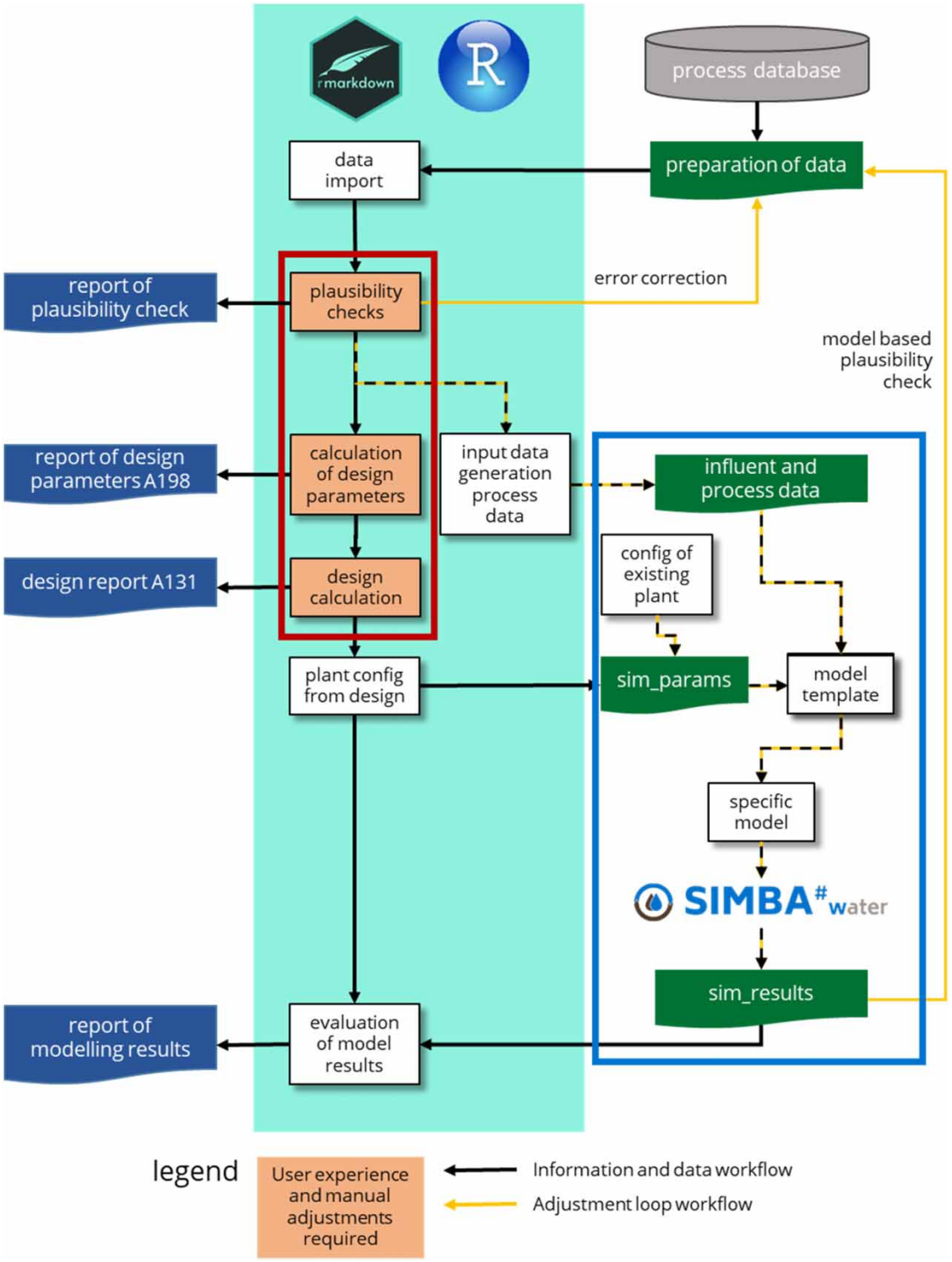


Figure 3 | Scheme for the sample workflow for the design and simulation of a wastewater treatment plant based upon the applicable rules and regulations.

Preparation and data import

The data available in a spreadsheet format was checked with regard to the data columns it contained. A general data structure with all potentially important data points was created for integration into the calculation workflow. By entering column numbers, a direct link was made with the raw data to be evaluated. These were then imported

into the data structure. This procedure also allows other evaluations to be carried out with the data. A repeated import is not necessary.

An automated linking via the data column names is also conceivable, but due to the variety of error sources and the comparatively small number of data columns used, this does not make sense. In addition, a manual review and evaluation of the available metadata should be provided as a first step. In this way, it can be checked whether the same time references are available for the mean value formations for the daily values. For example, practice shows that 24-hour composite samples are taken from 7 a.m. to 7 a.m. to determine the influent concentration, while the influent volume flow is averaged from 0 to 12 p.m. This error must be excluded in advance in order to be able to determine correct influent loads.

The import script automatically detects data gaps, e.g. in the case of missing daily data records, and assigns the data correctly in terms of time.

Plausibility checks

This processing step is the essential basis for all further calculation modules. Due to the complexity of possible test methods and evaluation options, only aspects relevant to workflow are discussed here. Literature on this topic specifically for WWTP data was not found. Some basic explanations can be found, for example in Newhart *et al.* (2019). In this article, the necessity of an initial inspection – i.e. a visual interpretation – is emphasised. The comparatively high manual effort is also emphasised. Newhart *et al.* (2019) consider the necessary expertise to be particularly important: 'Identifying the structure and characteristics of a dataset requires familiarity with the source of the data and the process itself' (page 502).

From our own experience, the work in such design processes is very much intuitive and based on knowledge obtained from previous projects. A systematic approach is rather rare. Figure 4 shows plots from an example application of this module in workflow. While figures and descriptions are generated automatically, inspection and evaluation of reliability is carried out manually.

In addition to a visual representation by means of time series, the use of scatterplots of related variables (e.g. different influent concentration parameters) is useful. The temporal course of ratio values can also indicate seasonal effects (e.g. BOD/COD ratio in the inflow with different pre-degradation during long flow times in the sewer network). The use of histograms and cumulative frequency plots provide information on distributions and fluctuation ranges. Descriptive statistics can also be used to obtain characteristic values for evaluating data quality. As a result of these checks, individual data points or entire daily data sets are identified that can be classified as unreliable.

Since these activities are time-consuming and are repeated for each new project, the automation of as many work steps as possible increases efficiency and avoids errors. It is also necessary for quality assurance. This corresponds to the path of transforming data into information described by Corominas *et al.* (2018) – literally the transformation of data graveyards into data mines. Due to the special conditions in the wastewater sector regarding scarcity of concentration data, the authors point out the difficulties of adapting technologies for data evaluation from other disciplines.

Calculation of design parameters (A198)

This worksheet (ATV-DVWK-A198E 2003) regulates the determination of data for the design of various WWTPs. For the design, among other things, the design-relevant inflow quantity and the mass flow must be determined. These data refer to dry weather inflow conditions. The classification of a daily influent flow rate as a dry weather value can be carried out by three methods:

- Use of the weather key documented in the operational data (numerical code from 1 to 7 for different weather situations).
- Use of recorded precipitation data
- Use of a statistical method (so-called 21 d moving minimum; all days are classified as dry weather days whose inflow is not more than 20% above the moving minimum of 10 days before and after).

The first two methods rely on the availability of additional data, the last method is applicable in all cases. The results from all three methods differ in most cases. Therefore, a purely automated processing of the workflow is not possible, as the user has to manually evaluate and select the (intermediate) results to be used.

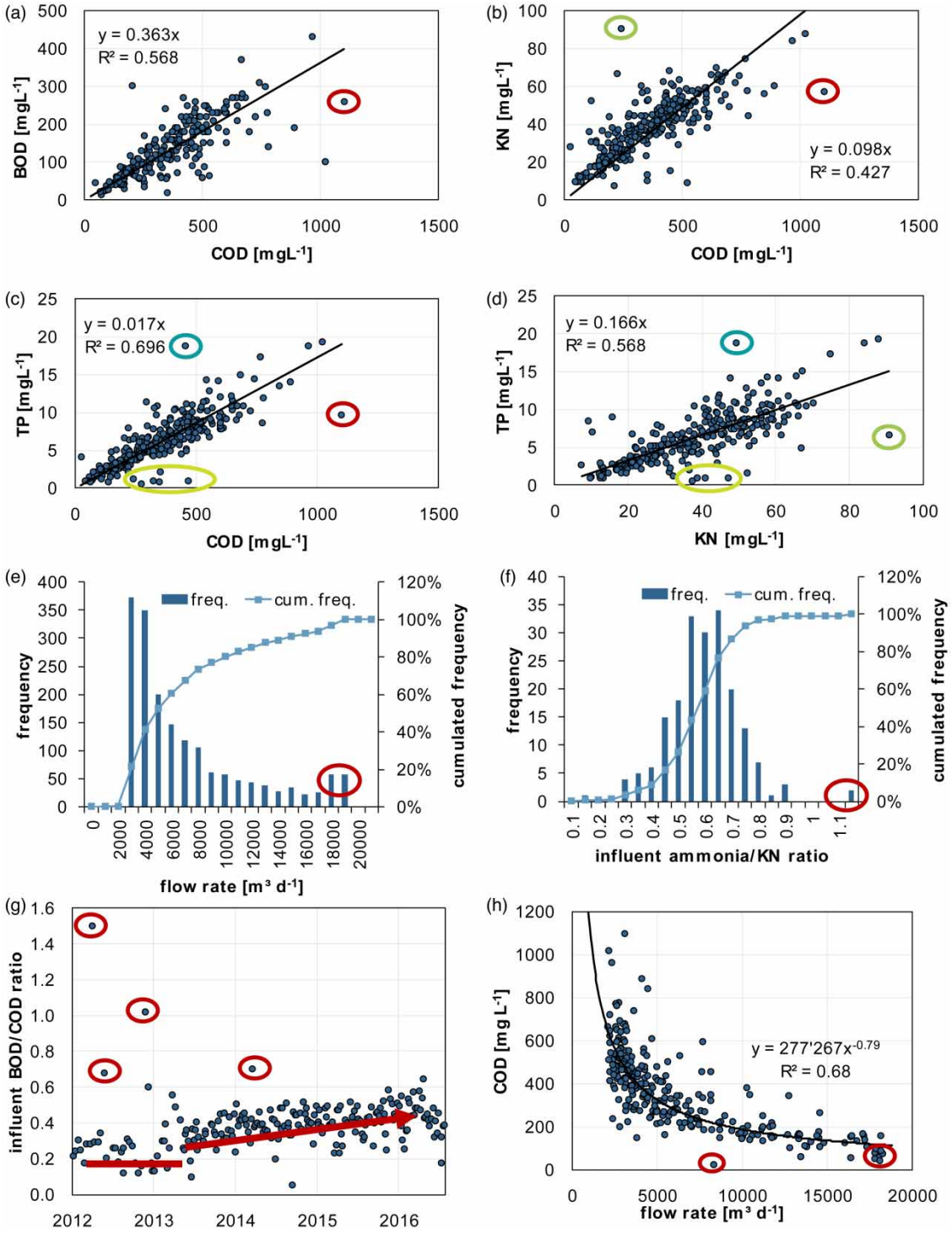


Figure 4 | Examples of plausibility checks: A-D Scatterplots of influent concentrations (marked unreliable ratios with colored circles); E/F histograms of influent flow rate and nitrogen ratio (marked anomalies); G time series plot of influent BOD/COD ratio with marked anomalies and increasing behaviour over time; H flow rate based check of influent COD concentrations (power function model as assumption of constant influent load) with marked unreliable value and possible influenced flow data from combined sewer overflow (bottom right).

Depending on the data density and subsequent treatment plant planning, the maximum 2- or 4-week mean values in the period with the design-relevant temperature (usually 12 °C) or the 85% percentile value are required to determine the influent loads.

During the development of the calculation script, the necessity of an exact definition of the quantities to be determined became apparent. For example, it remains unclear how to deal with the beginning or end of the period to be evaluated in the case of moving averages. It is obvious that these periods are not taken into account. This must be described in a subsequent documentation of the results.

When determining the design-relevant parameters in this module, manual intervention is necessary at various points. For example this results from the different methods for dry weather day determination and depends upon the available data. The manual selection method is described in the documentation. The transfer of the selected parameters into the calculation script is done with a separate parameter file, as it is shown in Figure 5. The main settings to be selected can be found to the right of the parameter input for the A 198.

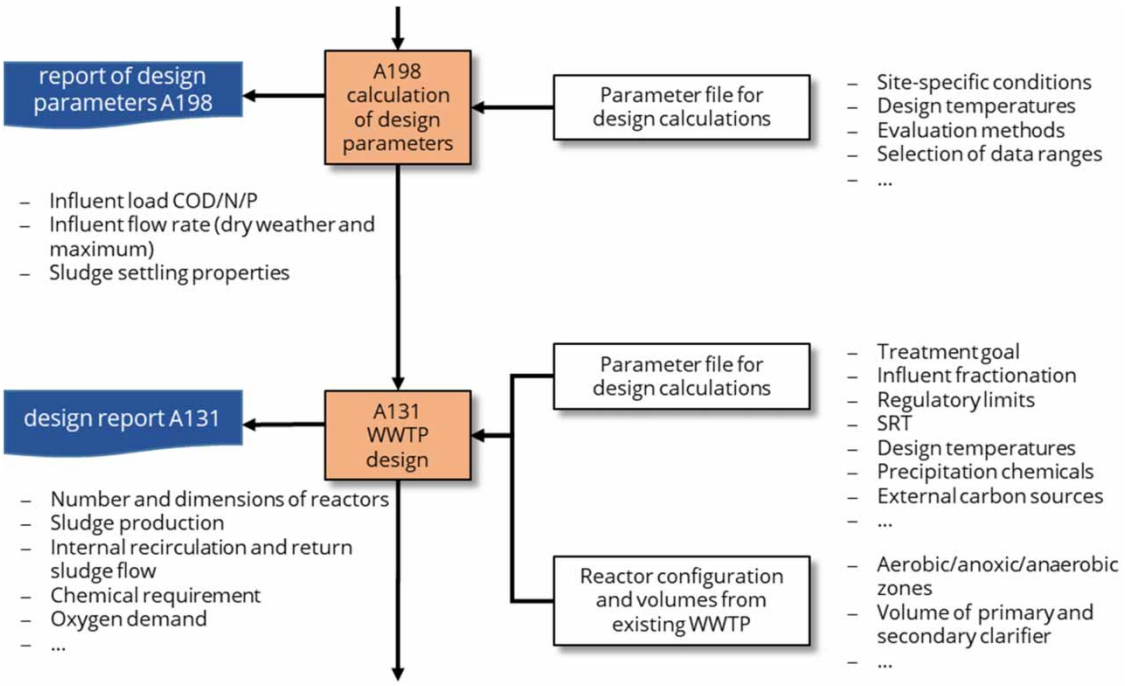


Figure 5 | Detail of the workflow: Input specifications and main results of the application of the regulations A198 and A131.

Design of WWTP (A131)

With the data compiled in the previous process steps and their aggregation to design parameters, a WWTP design is carried out in this module (based on the calculation specifications in DWA-Topics_T4/2016 2019). The calculation algorithm is basically linear. In an iteration cycle, the fraction of the denitrification volume (modified Ludzack-Ettinger process) is determined as a function of the oxygen recovery through the denitrified nitrate.

In contrast to the calculations with the A198, it is clearly and unambiguously specified here within which limits input parameters are to be selected manually and how the calculation process is carried out. These parameters are specified in a separate script document in order to maintain transparency (Figure 5). As a result, the tank volumes required for denitrification, nitrification and secondary sedimentation are calculated as well as the required oxygen input quantities and the internal recirculation as listed below the A131 module in Figure 5. This means that the basic parameters for the design of the plant are available and can be integrated into a BIM process.

The design calculations and the results are documented in a standardised report.

Activated sludge modelling

Activated sludge modelling is a proven tool in WWTP planning processes. Available instructions (e.g. Rieger *et al.* 2012) enable systematic processing. Alex *et al.* (2007) adapted ASM 3 (Gujer *et al.* 1999) to the DWA regulations. A static simulation yields results of effluent concentrations similar to the design results of DWA design guideline

A 131 (DWA-A131 2016. Thus, dynamic simulation can be used as a supplement to static design for a wide variety of tasks.

Automated integration into a process sequence is a sensible alternative to the usual procedure in order to minimise the workload and avoid transmission errors.

In the process diagram (Figure 3) this is shown as a separate block next to the R-based area. The WWTP models of the simulation programme Simba# (ifak 2021) can be fully parameterised via interfaces to spreadsheet programmes. All settings in the model (e.g. geometric data, but also technical and kinetic parameters of the activated sludge model, influent fractionation) can be adjusted and are simultaneously documented to provide transparency.

The input data of the model are based on existing operating data, which were prepared in previous process steps of the R-based modules. They can thus be easily and automatically transferred to the simulation environment. Thus users are able to dynamically simulate the existing plant with the available operating data. This can be used as an additional plausibility check (yellow arrows in Figure 3).

Furthermore, the newly planned WWTP can be dynamically simulated with the available load data and thus promote further information for detailed planning of technical equipment for example. It is helpful to distribute the hourly air volume demand for the correct technical design of the blower station. In addition, the model can serve both as a basis for detailed planning and for planning the commissioning and later for optimisations during operation, i.e. tasks that are summarised under the term 'digital twin'.

If the WWTP has a rather simple structure (e.g. classic modified Ludzack-Ettinger process), an already created model template can be parameterised appropriately and does not have to be newly created.

DISCUSSION

Practical advantages of the developed methodology

The workflow was developed in modules. Some of these modules have already been in practical use for some time. The following advantages can be derived from this. They can be categorised as 'time saving' and 'quality increasing' and are summarised in Table 1.

Table 1 | Advantages of the described digital workflow

Time saving through:	Increase in quality through:
Almost no manual data processing necessary	Traceability
Reuse in other projects	Transparency of the work steps
Simple corrections, extensions and additions possible	Documentation of the methodology
Essential parts of the documentation are prepared automatically	Absence (minimisation) of errors in the results
Iterative application easily possible	Uniform representations and documents
Better cost estimation for subsequent process steps	Focus on engineering issues due to time savings
Structured and standard-compliant procedure	

In addition, the time savings reduce project costs. Further positive influences on the cost structure arise from the faster and more concrete project planning capability, because much of the information required for this can be determined very quickly.

Lessons learned from development process and first use

The usual workflow, especially in smaller engineering consulting companies, corresponds to the procedure described in the left-hand column of Table 2. Although these are recurring tasks, the application of template documents, calculation tables or the use of self-programmed evaluation algorithms is rather low. Each new project is largely started from scratch. This means a highly unnecessary expenditure of time, which is further increased by errors in the application of the design specifications in newly created calculation documents.

The steps shown in the right column largely eliminate these disadvantages. Only once (and due to changes in the rules and regulations) does a higher effort have to be invested in the creation and testing of the digital workflow.

Table 2 | Comparison of atual typical and proposed workflow design

Production step	Usual workflow	Digital workflow
Pre-processing of data	Provide digital availability of the data	
	– Conversion of the data into necessary formats – Plausibility check of the raw data (manual or with simple template files), cleaning if necessary – Generation of a new corrected data table – Documentation with manual integration or links to the data table and evaluations	– Import and automatic data conversion – Standardised plausibility check with manually supported correction – Preparation of data tables for further evaluation – Automatic report generation
Data evaluation	– Analysis of the necessary calculation steps from the rules and regulations for data evaluation. – Implementation of the calculations, usually in a spreadsheet programme – Documentation through manual report generation with insertion or links to the data and evaluations (figures, tables)	– One-time definition of the algorithm in the workflow based on the set of rules – Reproducibility of the calculations through the created code – Automatic report generation of input data and design results with all tables and figures
Post-processing	– Transfer of the design results via report documents – Usually no standardised dissemination interface	– Digital transfer of input data, physical and chemical conditions and design results possible without interruption – Preparation of required interfaces necessary – Automated documentation of the interface possible

From the programming implementation and the first use cases, the aspects explained below have emerged as particularly important:

Time saving usable for engineering penetration instead of recurring unnecessary manual work with data

By using already existing submodules (data import and plausibility check), a detailed insight into the quality and density of the existing operational data can be obtained in the shortest possible time for a new WWTP. This enables a precise assessment of the possibilities with the existing data and the possible need to collect additional information.

Clear calculation specifications required

The implementation of calculation processes in programme scripts is only possible on the basis of precisely available algorithms. However, as soon as data are summarised over time intervals (weekly or monthly averages) or sliding calculations (average, minimum or maximum values) there are often uncertainties as to how these are to be realised. This concerns, for example, the start and end range in sliding calculations. The method chosen in the script must be indicated in the documentation of results.

No complete automation

A complete automation of the described workflow is feasible from a computational point of view, but from an engineering perspective it does not make sense and may even involve risks. The user experience is an essential tool in the assessment of the available data. However, decision-making can be considerably accelerated and objectified by technical support.

When determining essential (intermediate) results of the calculations, a manual specification of the values to be chosen is the most practicable solution. This can be integrated into the parameter scripts that contain the settings and defaults (see the next point).

Separation of calculations and defaults/parameters

Typical evaluations and measurements of a WWTP are based upon a large number of settings and characteristic values that are used in the individual calculation steps. In order to maintain an overview and simultaneously

make it easier to use the calculation scripts, it is recommended to separate them into calculation and input scripts. Corresponding documentation through comments supports usability.

Use of the documentation options

The choice of the programming language R in combination with document creation based on RMarkdown was made to document the results of the calculations consistently and transparently from within the calculation workflow. The direct coupling between calculation and documentation saves considerable processing time and minimises transmission errors. This can also be realised to a limited extent with office software, but with much reduced transparency and traceability.

The documentation is used for the presentation of results, the description of methods, and for basic explanations.

Subsequent editing of the generated documents is also conceivable. In this case, the direct link to the calculations is lost. When the script is executed again, the manual additions are lost or have to be repeated. Therefore, it is recommended to store supplementary explanations in a separate document with reference to the automatically generated report.

Preparing or thinking ahead for extension or reuse

Evaluation and assessment methods are subject to change. In addition, new and further developments based on previous information are conceivable. Therefore, calculation tools should be designed modularly and openly in order to enable future reuse with little effort.

Keep it simple

The selection of open-source scripting languages is based on the use of the existing basic functionalities and special packages or libraries based on them for specific requirements, e.g. for visualisation or statistical evaluation. The number of sources used should be minimised in order to maintain operability even when updating of certain specific packages is discontinued.

This requirement also applies to the way the code is created, including sufficient documentation and sensible variable nomenclature based on the names in the calculation specifications.

Disclosure of calculations

The specifications in regulations form the basis for the design and construction of wastewater treatment plants. These regulations describe the calculation methods to be used in varying degrees of detail. The programming implementation is associated with financial cost, which can be saved by other users if the code is made available as open source.

It is conceivable to integrate the creation of the calculation scripts into the development process of the rules and regulations – as a fundamental component of the respective rules and regulations. These can then be kept up to date and made available to all users via freely available platforms for software version management and distribution (e.g. https://github.com/). For this to happen there must be a willingness to share scripts. Collaborative development similar to open source software development is not yet to be expected in this field, at least not in German-speaking countries. Due to the large number of small consultants in the market there is little competition because many operators work with the same consulting firms for a long time. Therefore, there is little need for the companies to develop competitive advantages through innovative methods.

CONCLUSIONS

This paper describes a methodology with which an engineering workflow for the evaluation and assessment of routine data from WWTP can be implemented digitally. The main advantages are time savings on repetitive tasks, quality improvement through minimisation of error sources, and transparency through standardised documentation. This means that project processing time can be better used for the development process.

The methodology described is also suitable for integration into a BIM process, which corresponds to a digital twin in the construction sector. For this and other applications, use as a web-based solution is required. Appropriate security precautions must be observed. As explained by Ooms (2013), the programming language R was originally developed for offline use on a local machine. Therefore, there are hardly any restrictions on the execution of the programme. The application in a shared environment must therefore detect and prevent dangerous behaviour of the code or unrestricted use of hardware resources. Various approaches exist for this, as can be read at Ooms (2013).

New challenges arise in the application of this method, which are based on the diverse areas of expertise listed in Figure 1. The requirement profile of the applying engineer becomes larger, since they require information technology knowledge in addition to the purely urban water management knowledge.

As a result of the application test modules have been developed in R for the following tasks:

- Data import via a column-based interface
- Standardised plotting of various diagram formats
- Performing plausibility checks and generating a report
- Data evaluation according to A198 and generation of a report
- Wastewater treatment plant design according to A131 and generation of a report
- Preparation of influent and operational data for model adjustment
- Parameterisation of a predefined model template
- Documentation of the simulation results

A first development status of the described workflow can be found in the repository https://github.com/margon0815/digit_wwtp. In the basic state, the essential modules are included with some demonstrative examples. In the further progress of the project, extensions and additions will be made.

DATA AVAILABILITY STATEMENT

All relevant data are available from an online repository or repositories https://github.com/margon0815/digit_wwtp.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Alex, J., Wichern, M., Halft, N., Spering, V., Ahnert, M., Frehmann, T., Hobus, I., Langergraber, G., Plattes, M., Winkler, S. & Woerner, D. 2007 *A Method to use Dynamic Simulation in Compliance to Stationary Design Rules to Refine WWTP Planning*. Vienna, Austria, pp. 125–128, 9–13 September 2007.

Allaire J, X. Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W. & Iannone, R. 2021 rmarkdown: Dynamic Documents for R., R package version 2.11. Available from: https://github.com/rstudio/rmarkdown.

ATV-DVWK-A198E 2003 *Standardisation and Derivation of Dimensioning Values of Wastewater Facilities – April 2003*. German Association for Water, Wastewater and Waste (DWA), Hennef.

Blair, G. S., Henrys, P., Leeson, A., Watkins, J., Eastoe, E., Jarvis, S. & Young, P. J. 2019 Data science of the natural environment: a research roadmap. *Frontiers in Environmental Science* **7**(121), 1–14.

Blischak, J. D., Carbonetto, P. & Stephens, M. 2019 Creating and sharing reproducible research code the workflowr way. *F1000Research* **8**, 1749–1749.

Conway, D. 2013 *The DataScience Venn Diagram*. Available from: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram.

Corominas, L., Garrido-Baserba, M., Villez, 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**, 89–103.

Di Biccari, C. & Heigener, D. 2018 *Semantic Modeling of Wastewater Treatment Plants Towards International Data Format Standards*. Weimar, Germany, pp. 183–190.

DWA 2018 *Topic T4/2016: Design of Wastewater Treatment Plants in Warm and Cold Climates*. German Association for Water, Wastewater and Waste (DWA), Hennef.

DWA-A131 2016 *Dimensioning of Single-Stage Activated Sludge Plants. German Standard DWA-A 131 (Bemessung von Einstufigen Belebungsanlagen – Juni 2016)*. German Association for Water, Wastewater and Waste (DWA), Hennef.

DWA-Topics_T4/2016 2019 *Design of Wastewater Treatment Plants in hot and Cold Climates (EXPOVAL) - Corrected Version May 2019*. German Association for Water, Wastewater and Waste (DWA), Hennef.

Ferrer, J., Seco, A., Serralta, J., Ribes, J., Manga, J., Asensi, E., Morenilla, J. J. & Llavador, F. 2008 DESASS: A software tool for designing, simulating and optimising WWTPs. *Environmental Modelling & Software* **23**(1), 19–26.

Gabaldón, C., Ferrer, J., Seco, A. & Marzal, P. 1998 A software for the integrated design of wastewater treatment plants. *Environmental Modelling & Software* **13**(1), 31–44.

Gibert, K., Horsburgh, J. S., Athanasiadis, I. N. & Holmes, G. 2018 Environmental data science. *Environmental Modelling & Software* **106**, 4–12.

Gujer, W., Henze, M., Mino, T. & van Loosdrecht, M. 1999 Activated sludge model no. 3. *Water Science And Technology* **39**(1), 183–193.

Henze, M., Gujer, W., Mino, T. & van Loosdrecht, M. 2000 *Activated sludge models ASM1, ASM2, ASM2d and ASM3*. IWA Scientific and Technical Report No.9. IWA Publishing, London, UK.

ifak 2021 *Simba Water#4.3*. ifak e.V. Magdeburg, Germany.

ISO_16739-1 2018 *Industry Foundation Classes (IFC) for Data Sharing in the Construction and Facility Management Industries – Part 1: Data Schema*. International Organization for Standardization.

ISO_29481-1 2016 *Building Information Models – Information Delivery Manual – Part 1: Methodology and Format*. International Organization for Standardization.

Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., Jiang, N. & Halpern, B. S. 2017 Our path to better science in less time using open data science tools. *Nature Ecology & Evolution* **1**(6), 0160.

Manig, N., Beier, M., Rosenwinkel, K.-H., 2019 In: *Urban Water Management for Future Cities: Technical and Institutional Aspects From Chinese and German Perspective* (Köster, S., Reese, M. & Zuo, J. e. eds.). Springer International Publishing, Cham, pp. 57–68.

Marzouk, M. & Othman, A. 2017 Modeling the performance of sustainable sanitation systems using building information modeling. *Journal of Cleaner Production* **141**, 1400–1410.

Newhart, K. B., Holloway, R. W., Hering, A. S. & Cath, T. Y. 2019 Data-driven performance analyses of wastewater treatment plants: a review. *Water Research* **157**, 498–513.

Ooms, J. 2013 The RAppArmor package: enforcing security policies in R using dynamic sandboxing on Linux. *Journal of Statistical Software* **55**(7), 1–34.

Qiu, Y., Li, J., Huang, X. & Shi, H. 2018 A feasible data-driven mining system to optimize wastewater treatment process design and operation. *Water* **10**(10), 1342.

R Core Team 2021 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from: https://www.R-project.org/.

Rieger, L., Gillot, S., Langergraber, G., Ohtsuki, T., Shaw, A., Takacs, I. & Winkler, S. 2012 *Guidelines for Using Activated Sludge Models*. IWA Publishing, London, UK.

Söbke, H., Theiler, M., Tauscher, E. & Smarsly, K. 2018 BIM-based description of wastewater treatment plants. In: *Proceedings of the 16th International Conference on Computing in Civil and Building Engineering (ICCCBE)*, Tampere, Finland, Vol. 6(05), p. 2018.

Söbke, H., Abadia, P. P., Heigener, D. & Smarsly, K. 2020 *BIM-based Sizing of Reactors in Processing Facilities*. Universitätsverlag der TU Berlin, Berlin, Germany, pp. 412–421.

Söbke, H., Peralta, P., Smarsly, K. & Armbruster, M. 2021 An IFC schema extension for BIM-based description of wastewater treatment plants. *Automation in Construction* **129**, 103777.

Sonaje, N. P. & Berlekar, N. D. 2015 Modeling of wastewater treatment plant design for pulp and paper industry: a review. *International Journal of Civil, Structural, Environmental and Infrastructure Engineering Research and Development (IJCSEIERD)* **5**, 59–68.

Stodden, V., Leisch, F. & Peng, R. D. 2014 *Implementing Reproducible Research*. Chapman and Hall/CRC Press, Boca Raton, London, New York.

US-EPA 2015 *Continuous Monitoring Data Sharing Strategy*. EP-C-12-052 Task Order No. 0005, US EPA, Washington, DC.

Xie, Y. 2018 *knitr: A Comprehensive Tool for Reproducible Research in R*. Implementing reproducible research. Chapman and Hall/CRC Press, Boca Raton, Florida, New York, pp. 3–31.

First received 6 April 2022; accepted in revised form 15 September 2022. Available online 23 September 2022