Interreg
Danube Region

Co-funded by
the European Union

Tethys

Output 1.5

Transnational hazardous substances database coupled with data-processing and pollution assessment package as operative tool to support regular transnational modelling-based status, risk and scenarios assessment

Solution jointly developed and implemented by:

TU Wien (AT), Budapest University of Technology and Economics (HU), International Commission for the Protection of the Danube River, Bulgarian Water Association (BG), National Administration „Romanian Waters" (RO), Water Research Institute (SK), Jozef Stefan Institute (SI), Center for Ecotoxicological Research Podgorica (ME), Ukrainian Hydrometeorological Institute State Service on Emergencies and National Academy of Science (UA), Croatian Waters (HR), Public Institution "Vode Srpske" (BA), Environment Agency Austria (AT), Jaroslav Černi Water Institute (RS)

Authors:

S. Kittlaus, S. Osten, N. Weber, A. Kovacs, T. Lajko, A. Clement, M. Kardos, K. Dudás, O. Zoboli, M. Sidau, I. Lupu, A. Daescu

December 2025

Table of content

Project co-funded by European Union funds Danube Region Program and National Funds of the participating countries

Abstract

Pathway-oriented management of hazardous substance (HS) pollution at the river basin scale requires harmonised, quality-assured concentration data from multiple environmental compartments to support emission modelling. In the Danube River Basin, such integrated data management tools were initially developed within the INTERREG Danube project *Danube Hazard m³c*, resulting in a transnational PostgreSQL database comprising 10.7 million measurements from over 383,000 samples at approximately 25,000 sites. Although highly valuable for assessing data availability and supporting emission modelling, the original database showed limitations in efficiency and long-term operability. Within the Tethys project, the database was comprehensively revised, extended and optimised through a pilot action, transforming it from an expert-oriented system into an operative management tool. The upgraded database features an improved structure, extensive quality control and assurance workflows, enhanced interoperability with EU reporting databases, and dedicated tools for data evaluation and processing. It provides quality-checked data and rich metadata for parameterisation and validation of river basin–scale emission models, while also supporting pollution status assessments and trend analyses. The ICPDR intends to adopt the database as an operational tool for transboundary HS pollution assessment and to explore its integration into existing water quality information systems.

1 Introduction

1.1 Purpose of the database and pollution evaluation tool

State-of-the-art management of hazardous substance (HS) pollution on the river basin scale needs a sound understanding of the extent and pathways of surface water pollution. This can be achieved by smartly coupling targeted monitoring and emissions models following the pathway-oriented approach. However, such models need not only river monitoring data to validate them but additionally substance-specific concentration data from different environmental compartments relevant to derive loads of associated emission pathways. If such data are collected for a wide range of relevant pollutants, proper data management with clear quality assurance procedures becomes necessary.

On the Danube basin scale, such tools are not available to support data collection with proper quality assessment routines not only for surface water data but for a wider spectrum of monitoring data. Within the INTERREG Danube project Danube Hazard m³c, a first transnational database for collection of hazardous substance concentrations was developed to support emission modelling. Data from multiple environmental compartments in many countries within the Danube river basin were collected, harmonized and stored in a PostgreSQL database. The results were remarkable, with the database containing 10.7 million concentration measurements in over 383 000 samples from approximately 25 000 sampling sites from the Danube countries. This database proved very valuable to assess, quantify and compare data availability in the Danube basin, to identify priorities for future monitoring programmes and it also allowed to derive essential input and validation data for emission modelling (Kittlaus et al., 2024)[1]. Nevertheless, the fact that most of the resources and efforts within the Danube Hazard m³c project had to be invested in the search, collection, harmonisation and evaluation of the data, after the project completion a critical evaluation of the technical tool itself revealed that some crucial aspects were suboptimal and hindered its sustainable use. Therefore, in the Tethys project a pilot action was carried out to jointly extend, upgrade and demonstrate the existing database to make it a more robust and sustainable instrument for the management and control of hazardous substances in the Danube River Basin.

The process of implementation, testing and demonstration of the pilot action as well as the technical details are described in the Tethys Output 1.3. This document describes the jointly developed solution, by summarising the conceptual framework and the main technical aspects of the database.

The main aim of the new tool is the supply of quality checked data and extensive related meta data for the parametrization and validation of emission modelling at the river basin scale. Aside from this main purpose, the tool can additionally contribute to a better understanding of the status of water pollution and support e.g. trend analyses and status assessment.

1.2 Ideas and principles leading to the implemented tools

As the requirement for better data management tools existed not only at the transnational level but also at national level in many Danube countries, it was decided to develop together the tools for national and transnational use. This has a big advantage for sharing data between national and

[1] Kittlaus, S., Kardos, M. K., Dudás, K. M., Weber, N., Clement, A., Petkova, S., Sukovic, D., Kučić Grgić, D., Kovacs, A., Kocman, D., Moldovan, C., Kirchner, M., Gabriel, O., Krampe, J., Zessner, M. and Zoboli, O. 2024 A harmonized Danube basin-wide multi-compartment concentration database to support inventories of micropollutant emissions to surface waters. *Environ Sci Eur*, **36**(1). https://doi.org/10.1186/s12302-024-00862-4

transnational database. Such data transfer will be necessary on the one hand for transferring national monitoring data to the transnational database as support for transnational emission modelling. On the other hand, it might be helpful for national model application, e.g. to more easily retrieve data from neighbouring countries from the transnational database.

To make the tools easily applicable for all interested partners, it was decided to base the tools on freely available software, which require no further investments for license fees.

The user interfaces should be accessible without advanced programming knowledge and be understandable without the need to study an extensive user manual. Documentation should be incorporated into the tools as far as possible.

As the majority of the countries in the Danube basin is part of the EU or fulfils voluntarily **EU reporting obligations**, the data structure and controlled vocabulary should be harmonized with those of the EU WISE and UWWTD reporting data systems to reduce effort for data preparation and support the collection of data for EU reporting with these tools.

To achieve a high data quality which facilitates data evaluation, the database should use **controlled vocabularies** for those columns where similar concepts are used repeatedly. Such controlled vocabularies develop over time, which needs a organised way of item registration and deprecation. Here the requirements defined in

- ISO 25964-1:2011 Information and documentation —Thesauri and interoperability with other vocabularies
- ISO 19135-1:2015 Procedures for item registration

were considered and implemented where possible.

As mistakes happen where data are created and managed, it was decided to explicitly define a very thorough **quality assessment workflow** within the database with different roles: One role importing and updating the data and one role for checking the quality and raising issues on how to improve the data quality.

All datasets in the database should clearly show from which data source they were derived and which license applies, in order to facilitate reuse of the data as easily as possible but also to transparently constrain reuse where needed.

As the development of the tools was a coding project of significant size with multiple persons working on the same projects, it was considered necessary to use professional code hosting, versioning and releases.

A development process was set up, which allows creation and update of databases on servers not directly accessible for the development team.

It was decided to develop the tools for national and transnational use in parallel, in order to keep data structures as compatible as possible and allow for swift data exchange among databases. Extensions for the use of national language were included as optional features.

2 The Transnational hazardous substances database

2.1 Institutional stewardship

The development of the database was closely coordinated with the ICPDR Secretariat and the Pollution Management Expert Group so that it is in line with the needs of the Danube countries and the existing ICPDR data management tools and procedures.

The ICPDR fully supports the initiative of the future adoption and use of the database that shall be an essential operational tool for collecting, processing, analysing and disclosing data on chemicals concentrations and for supporting periodic modelling activities on the basin-wide level. In line with its transboundary coordination role, the ICPDR is willing to facilitate future potential updates of the database, closely linked to the updates of the national databases to be carried out by the Danube countries. In addition, the ICPDR will consider the possibility of integrating certain parts of the transnational database into the ICPDR water quality database. Moreover, the ICPDR intends to support the dissemination of relevant data to the interested stakeholders through its existing platforms.

Although the ICPDR represents the key beneficiary of the transnational database, its hosting, the technical maintenance and potential upgrade shall remain with one of the scientific institutes of the Danube countries, ideally closely linked to the maintenance of the basin-wide modelling tool. The systematic update of the database shall provide inputs for the periodic reporting on pressure assessments coordinated by the ICPDR (e.g. the 6-year cycles of the Water Framework Directive). Data updates shall be managed by the future modellers to be tasked by the ICPDR in close cooperation with the database hosting institute. On the one hand, various data can be directly migrated from the updated national databases maintained by the Danube countries, on the other hand, certain data shall be centrally collected and updated by the modellers. The overall data collection process shall be facilitated by the ICPDR through its professional network and national experts. This can ensure full transparency and acceptance: either official national data will be collected from the countries or in case of lack of national data, publicly available international data will be used that are accepted by the countries.

This process shall be jointly implemented based on agreement and request of the Danube countries, where the ICPDR shall ensure overall coordination and organizational support, while the hosting institute and the modellers shall perform the necessary technical work as appropriate.

The ICPDR will consider the possibility of integrating certain parts of the database into the Danube River Basin Water Quality Database so that complementary information on HS pollution to the regular monitoring data reported by the Danube countries can be provided. The technical implementation of any agreed data exchange shall be jointly carried out by the hosting institute and the ICPDR.

On top of this, the database shall be made available by the hosting institute for the stakeholders, scientific community and even for the wider interested public, while respecting all data confidentiality conditions. The ICPDR is willing to provide a thematic webpage on the Tethys project results with a link to the database.

In line with these ambitions, the ICPDR will organize a special technical session with the project representatives in spring 2026 to discuss the details and necessary steps of the adoption and future use of the project results at the ICPDR level, including the database.

2.2 Technical implementation

2.2.1 Concept

To allow for an intuitive user interface for persons working in water management on the one hand and to enable several advanced technical functionalities, the database is built with two layers, implemented in different database schema:

1. The **technical layer** holds the actual data and is only visible to administrators. This layer consists of tables all placed in the schema "hidden". Data from different environmental compartments are combined in few tables based on common data structures. Data which were deleted by the user are still available here and can be restored in case of unintentional deletion.
2. The **user accessible layer** presents the data in a more straightforward way to users. Here a different schema for each environmental compartment presents the data for this compartment only. From a technical point of view, all data presentations are only views of the data in the schema "hidden", presenting those columns and rows relevant for each specific environmental compartment. Deleted data are not shown.

2.2.2 Included data

The aim of the database is to be able to host all concentration measurements obtained through monitoring from different environmental compartments which are needed to generate input data and validation data for emission modelling, e.g. with the regionalized pathway approach (e.g. MoRE model), but which can also be of great value for other spatial and temporal analyses of contamination across environmental compartments.

To make the database also useful to manage data which needs to be reported under the EU WISE system, additional attributes were added to enable storing concentration measurements in biota and in sediments. Information about the WISE data format was derived from the WISE- SoE Data dictionary[2].

Only disaggregated data (single values) can be stored in the database, as these can support different approaches and aims of data evaluation much better than aggregated data.

The environmental compartments included now (future extensions are possible) and depicted in different schemata are:

- **Surface water**: Data from rivers, lakes and reservoirs including concentrations in the liquid phase, suspended particulate matter and biota (for WISE compatibility).
- **Groundwater**: Concentrations measured in springs and wells.
- Wastewater: Concentrations in municipal and industrial wastewater (treated and untreated) and in different kinds of sewage sludge.
- **Stormwater runoff**: Concentrations in stormwater runoff from urban surfaces sampled from storm sewers in the separate sewer system and combined sewer overflows (CSO) in the combined sewer system. Untreated wastewater sampled during storm events from combined sewers can be stored here or in the wastewater section. Untreated wastewater from dry weather periods shall be stored in the wastewater section.
- **Atmospheric deposition**: Deposition from the atmosphere onto surfaces. Given either as deposition rates (mass per area and time) or as concentration in a collection container which

[2] WISE- SoE Data dictionary: Dataset specification for WISE SoE - Water Quality in Inland, Coastal and Marine waters (WISE-6) *Version December 2023*, created 17/01/2024, European Environmental Agency (EEA)

was collecting deposition for a certain period. To calculate deposition rates from these concentrations, the exposed collection area, the exposition time and the total sample volume need to be known.

- **Soil**: Substance content in soil samples (usually topsoil) given in mass of a substance per dry matter mass of soil.
- **Sediment**: Substance content in river bottom sediment samples.

2.2.3 Structure

2.2.3.1 Schemata

The database is structured into different schemata, which hold the tables for the different environmental compartments and some schemata for general information.

Table 1: Database schemata.

Schema name	Description
atdep	Atmospheric deposition: Measurements of substance transfer from atmosphere to surfaces.
data	Data sources: Documentation of the data sources, owners, suppliers and licenses.
determ	Determinants: Definition of determinants (substances and other parameters like pH) their classification and their environmental quality standards (EQS).
groundw	Groundwater: Concentrations in water below the earth surface.
hidden	Schema containing all the data table itself. Only accessible for administrators.
public	The default schema created by PostgreSQL where all users have certain rights. If objects are created without assigning them a dedicated schema, they will be stored here. In this schema the quality checked data are presented for download.
sed	Sediments: Concentrations in solid layers below the water column.
soil	Soil: Concentrations in samples from (top)soil contributing to surface water pollution via soil erosion and via washout to groundwater.
stormw	Stormwater runoff: Concentration in runoff generated mainly by precipitation events. Either as runoff from specific urban surfaces, as runoff in the storm sewer in separate sewer systems or as combined sewer overflow (CSO) in the combined sewer system.
surfw	Surface water: Concentrations measured in the liquid phase, biota or suspended particulate matter in surface waters: rivers, lakes, reservoirs, coastal waters ...
wastew	Wastewater: Wastewater from municipal or industrial sources, treated or untreated. Including also concentration in sludge resulting from wastewater treatment.

All data are stored in tables in the schema *hidden*. For user interaction they are presented in the different schemata via database views.

2.2.3.2 Views in schemata besides schema *public*

The same data are presented in different views for different purpose:

- For the data **import** purpose, only the columns which can be filled during data import are shown. Additional columns filled automatically during data import are hidden (e.g. the numerical id, the time of import and the user importing the data). Here only those data created by the user itself are visible. Moreover, data are only visible in this view as long as they have not gone through the quality assessment workflow. These views are appended with the suffix *_import*.
- A different view of the data is available for the quality assessors in order to perform the **quality assessment**. Here some additional data are included (e.g. the user who created the data set) and only the two columns regarding quality assessment (*quality_check_passed,*

comments_quality_assessment) can be updated. These views are appended with the suffix *_qa*.

- The data which were **not accepted by the quality assessor** are presented in a further view to the user who imported the data for correction or deletion (and reimport). These views are appended with the suffix *_qa_rejected*.
- To check all imported data independently of their quality assessment status, views without a suffix are available for every table. These views are read-only. For changing or deleting data, the aforementioned views must be used.

At the top of every schema there are three support views whose names start with an underscore.

- The *_table_definition* view contains information about all views in a schema and should therefore enhance understanding of the database structure.
- The *_column_definition* view contains comprehensive information about all columns regarding their description, data type, constraints or keys and should therefore serve as a tool to support data import.
- The *_controlled_vocabularies* view contains predefined column entries for character columns where only selected entries might be used. They are originally stored in schema *hidden* and maintained by the content administrator role *tethys_ca*. They serve as foreign key for *_import* views and are marked as such in the *_column_definition* view.

2.2.3.3 Views in public

- Views for data evaluation and extraction: they bear the same name of the environmental compartments from which their data originates and contain quality checked data that areas merged into wide tables displaying all collected information. They only miss columns which are recorded for data traceability, e.g. columns with the originally reported text strings, which were mapped on the controlled vocabularies.
- Views to support the quality assessment workflow: The two views *_qa_awaiting* and *_rejected_data* provide an overview of data that either have not been quality assessed or have not been corrected yet. They aggregate information about data recently imported or rejected during quality assessment with their schema, viewname, number of rows and the name of the user who imported them and/or the quality assessor who quality checked them.

2.2.4 User rights management

User rights for working with the database depend on PostgreSQL roles. The following roles were defined:

user (e.g. athompson)

Every user gets a login name and password. With this account, the user can connect to the database. Technically speaking this is the session-user, which is recorded when creating, changing or deleting data. This role has no further privileges nor rights. To see, create, modify and delete data, the user needs to take an additional role by running the command SET ROLE tethys_... ;.

tethys_reader

The **reader** role allows to see and download the quality-checked database content. It has no privileges to create, modify or delete data. Having this role, the user can check the database without being able to do any harm to the data.

tethys_user

The tethys_user inherits all rights from the tethys_reader. Additionally, this **user** role allows to create new data and modify/delete these data created by the user itself. Once the data are quality checked by a quality assessor, the user cannot modify or delete the own data anymore.

tethys_qa

The tethys_qa role inherits all rights from the tethys_reader role. Additionally, the **quality assessor** can see data created by other users. It can comment on these data and mark them as quality checked or reject them and send them back for revision.

tethys_ca

The tethys_ca inherits all rights from the tethys_user role. Additionally, the **content admin** has the right to see, modify and delete all data in the database. In addition to the data visible for other roles, it can see and modify the tables in the schema "hidden", where all data are stored. Thus, it can change controlled vocabulary and restore accidentally deleted data.

tethys_admin

The tethys_admin inherits all rights from the tethys_ca role. Additionally, the **admin** role can create new users, grant roles to users and change the structure of the database.

2.2.5 Data quality assurance

The following procedures are implemented to ascertain data quality in the database:

- Traceability: Every record is automatically amended with the information of who created it (column *created_by*) and at what time (column *created_at*) as well as of who last modified it (column *updated_by*) and at what time (column *updated_by*).
  To document how pre-existing data were mapped on the controlled vocabulary of the database, for several attributes there is an additional text-column, in which the original value should be stored. This column has the suffix *_reported*.
- Deleted data from the tables are not deleted when user delete them but moved into a separate data structure hidden from the ordinary user ("soft-deletion") and available to be restored by the administrators on request.

The following constraints are implemented:

- Primary key constraints: They ensure that uniqueness as well as a not-null condition of a column entry or a combination of column entries are satisfied.
- Not Null constraints: They ensure that a not-null condition of a column entry or a combination of column entries are satisfied. In the *_column_definition* view, those columns are marked in the mandatory column.

- Check constraint: It ensures the satisfaction of certain conditions based on logical queries or the existence of other column entries in a row.

Example 1: Check constraints for numeric values
Wherever possible numeric values are checked on insertion for their plausibility (e.g. areas, discharges, inhabitants) to be within a meaningful range, usually > 0 and smaller a maximum value. The catchment area of a river must be larger than 0 km² and smaller than 7000000 km² (size of the Amazon river basin).

Example 2: Values are checked based on values in other columns
If a measurement is marked as below an analytical limit, the value of the limit must be given in the corresponding column. If a measurement is marked as below the analytical limit of detection (LOD), the LOD must be given in the corresponding column.

- Unique key constraint: Ensures that uniqueness of a column entry or a combination of column entries are satisfied. Some unique keys are supposed to store aggregated information about its row. If this is the case, it can be seen in the description of the column. In the *column_definition* view those columns are marked in the *col_constraints* column with (u).

Example 1: A combination of columns are used as a unique key
A measurement is defined as a unique combination of an environmental compartment, a determinant (substance), the analytical method used to measure, the laboratory where the analysis was conducted and the identifier of the sample. No further data row with the same combination can be imported, which avoids duplicated data in the database.

- Foreign key constraint: Ensures the case sensitive referencing of a column or a combination of column entries in the so called "child table" originally migrated from the "parent table", hence providing referential integrity. In the *column_definition* view those columns are marked in the *col_constraints* column with (f).
- Controlled vocabulary: Those are a special version of foreign keys, here used for qualitative attributes: Whenever a qualitative attribute is recorded, controlled vocabularies (a list of allowed values for that column) are used. These assure, that the same thing is always called the same (and without typos) and no further data cleaning is necessary before data evaluation. Controlled vocabularies are stored in separate tables. These tables cannot be changed by the ordinary user, but only by the *content_admin* role. In the *column_definition* view those columns are marked in the *controlled_vocabulary* column. Possible entries for those columns can be seen in the *_controlled_vocabularies* view.
- Quality assessment workflow: As described below.

2.3 Data management workflow

After a user (role tethys_user) uploaded data, they are shown in the import-view until they have gone through the quality assessment procedure. As long as they are shown here, they can be updated or deleted by the user. 30 minutes after the last modification of a data row, it becomes available in the quality assessment workflow: it is shown to the quality assessors in the quality-assessment views.

The quality assessor can now check the data, decide if the data and meta data are of sufficient quality or if further updates are needed. If the data are fine, the quality assessor marks them as quality checked. If the data needs revision, the quality assessor fills his remarks in the column *comments_quality_assessment* and sets the column *quality_check_passed* to FALSE.

The data needing revision are presented to the user who created them in a dedicated view (suffix _qa_rejected) where they can be updated or deleted. The quality assessor still sees them in the quality assessment workflow as long as they are not changed. When changes occur, they are not visible to the quality assessor until 30 minutes after the last modification, when they occur again.

The 30 minutes embargo time data are not shown to the quality assessor after creation and modification should avoid, that the quality assessor already process data, which are still checked or modified by the creating user.

The quality assessor itself cannot change the data. He only provides feedback to the user who created the data. While the roles with their rights a strictly separated, it is possible that the same persons can select different roles for different tasks.

The data assessed by the quality assessor and marked as checked occur in the *public* schema in views which merge all information about sampling sites, samples and measurements into wide tables which are easy to use. All users with access to the database have read access to this view, which enables them to download the high-quality data sets for evaluation and further usage.

2.4 Content of the database

The database contains multi compartment data sets which are either of international, national or regional origin and which were collected through monitoring programmes ranging from official surveillance to research-based studies. The data origin is indicated by the data source identifier, which must be filled in every table related to the data set.

Beside those, there are data source unrelated tables listing determinants, analytical methods, etc. or in general the afore mentioned controlled vocabulary.

In the course of the Tethys project, every partner imported at least the data that have been generated during the Tethys project monitoring activities (Tethys Outputs 1.1 and 1.7) and the associated meta data, but some partners imported additional large data sets of high relevance.

The current status of the most important tables is shown in Figure 1, which shows the compartment specific amount of data in the three most important data source related tables. Figure 2 shows the number of measurements per country, while Figure 3 shows the share of Tethys project measurements per country. This amount of data largely fulfils the main objective of this activity in the Tethys project, which was primarily the technical development and demonstration of the database as instrument to support transnational control and management of HS in the Danube region. Further data will follow with more time. Most partners are in the process of implementing fully compatible national databases and of importing their additional data sets until the project completion, which can then be transferred to the transnational database with minimal effort. The increasing share of data collected in monitoring campaigns other than the Tethys project's one imported in the database, represented by the grey areas in Figure 1 and Figure 3, shows a promising and positive trend of use by several partners.

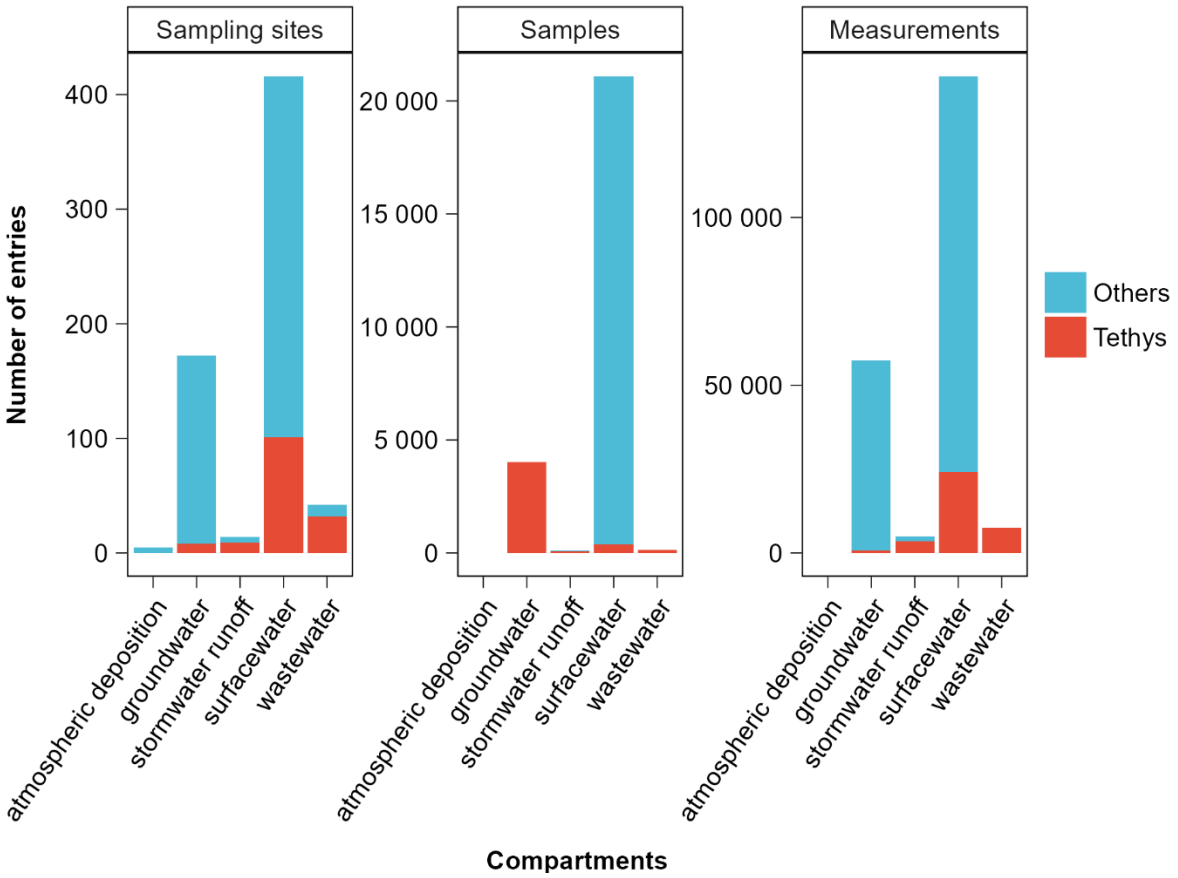

Figure 1: Overview of the amount of data included in the transnational database at document completion. Color indicates the data source. Tethys: monitoring within the Tethys projects. Others: monitoring from other sources.

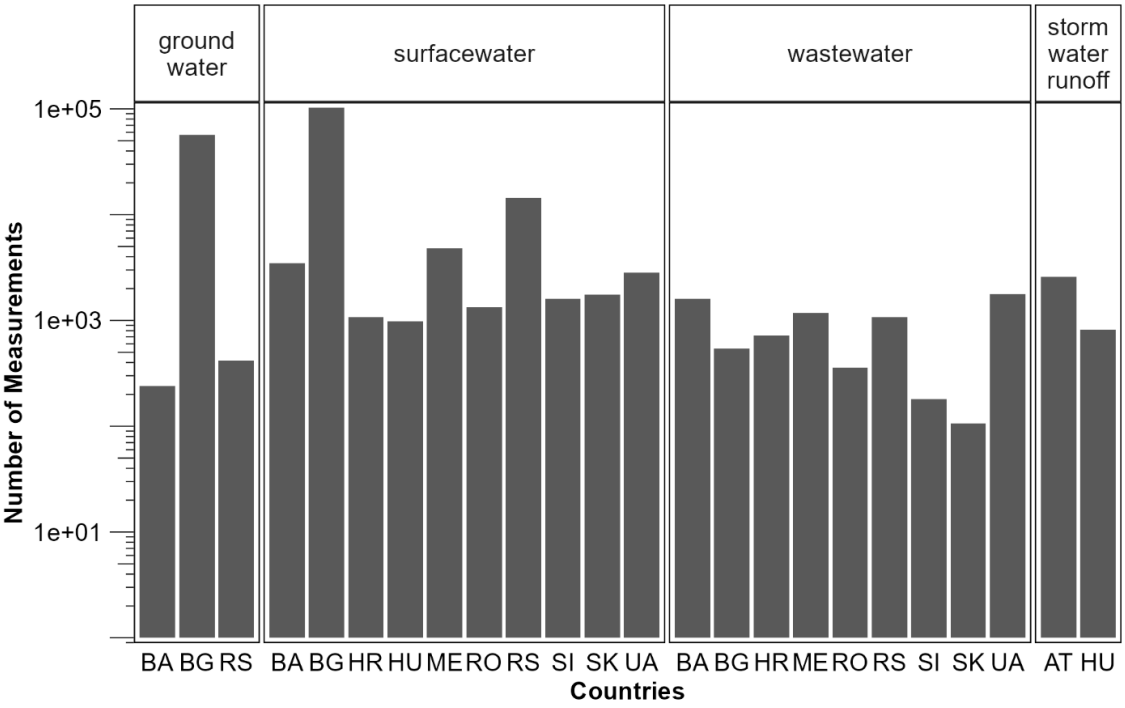

Figure 2: Number of measurements per country in the transnational database at document completion.

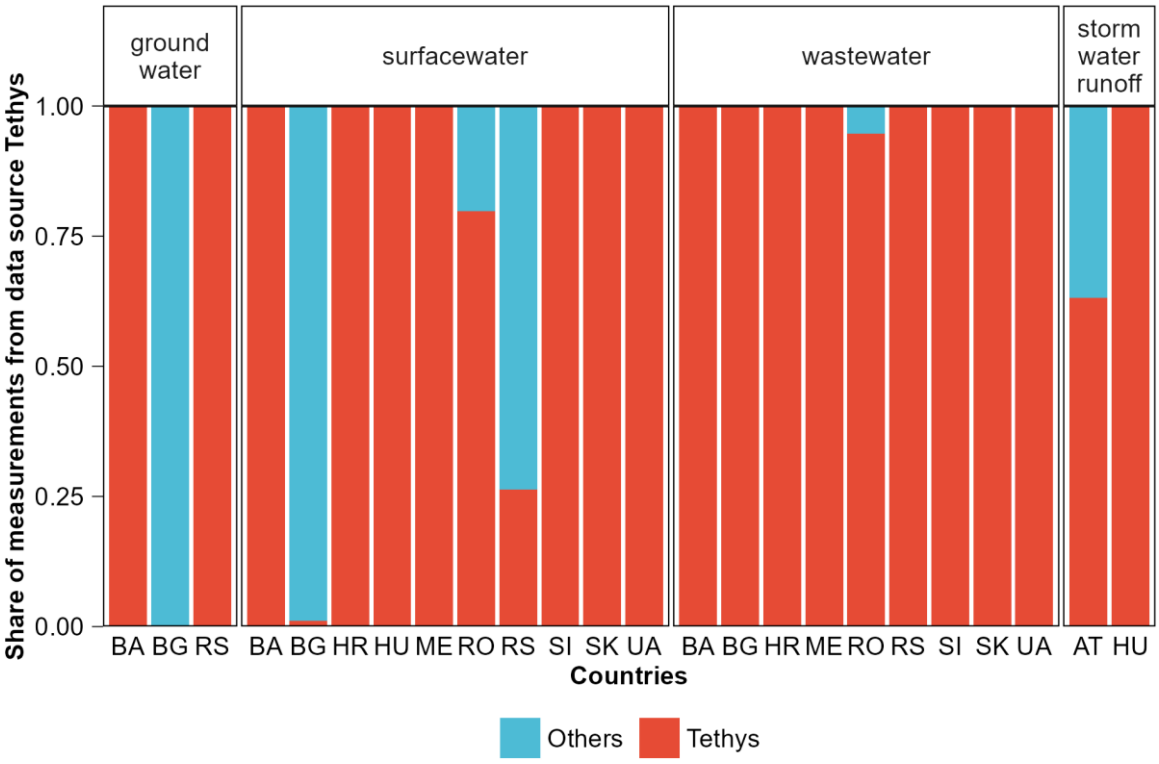

Figure 3: Share of measurements from the Tethys monitoring and from other sources within the transnational database at document completion.

2.5 Use of the database to support transnational emission modelling in the project

The transnational database has been used for two purposes: (a) to derive the emission factors for the municipal (and to a lesser extent, for the industrial) wastewater treatment plants and (b) to derive the validation data (immission concentrations and loads).

2.5.1 Emission factors for municipal WWTP

When developing emission factors (EF), data can be structured according to multiple classification dimensions. The most straightforward dimension is treatment technology, which categorizes facilities by their specific treatment processes (i.e., primary, secondary, tertiary, or quaternary). This classification enables direct application of emission factors when modelling future treatment scenarios. However, this approach may obscure geographically driven variations in pollutant concentrations.

An alternative classification dimension is wastewater treatment plant (WWTP) size, defined by design capacity (typically expressed in population equivalents). This approach was successfully applied to per- and polyfluoroalkyl substances (PFAS) in the PROMISCES project. However, for certain substance-region combinations, the available dataset may be insufficient to generate statistically robust emission factors. Additionally, establishing defensible capacity thresholds for size categories requires careful consideration of data distribution and statistical representativeness.

Geographic location represents another critical classification dimension, particularly for PFAS and pharmaceutical compounds where industrial discharge patterns, consumption habits, and regulatory frameworks vary substantially across regions. This report presents selected findings that illustrate these various classification dimensions, though comprehensive coverage results from data availability rather than systematic design.

The fundamental methodology employs measured effluent concentrations as emission factors. Total pollutant mass discharge from a treatment plant is calculated using the equation E = Q × C, where E represents the mass emission rate [kg/year or g/day], Q represents the volumetric discharge rate [m^3/year or m^3/day] for the specified time period, and C represents the emission factor [mg/L or µg/L] specific to the substance, treatment configuration, and potentially additional parameters such as geographic region or plant capacity class.

Alternative normalization approaches based on population equivalents (PE) or per capita (plant-connected inhabitant) values were also investigated and evaluated but are only partly presented herein. Population equivalents in wastewater engineering are standardized against biochemical oxygen demand (BOD_5 = 60 g/PE/day), which effectively normalizes for organic loading. This normalization methodology systematically masks the contribution from industrial discharges, which can dominate micropollutant concentrations independently of total organic load. While the actual connected population (inhabitants) would provide a more accurate basis for estimating domestic pollutant contributions, these data are unavailable in the UWWTD database and similar multinational databases.

In the following, we represent the approaches to deliver EF. The final emission factor values are provided in detail in the Tethys Outputs 2.1. and 2.3.

2.5.1.1 Metals

Two of the above-mentioned approaches used in case of metals.

Copper, zinc, arsenic, cadmium and lead

A widely implemented technique for enhanced phosphorus removal at wastewater treatment plants employing tertiary treatment processes involves the addition of aluminum-based salts (such as aluminum sulfate or polyaluminum chloride) or iron-based salts (such as ferric chloride or ferric sulfate) to the wastewater stream. These coagulant chemicals react with dissolved phosphate species to form solid precipitates and promote the formation of larger aggregated structures known as coagulates or flocs within the wastewater matrix. Phosphorus compounds, which under typical wastewater conditions would otherwise remain in dissolved or colloidal form in the aqueous phase, exhibit a strong tendency to adsorb onto the extensive surface area provided by these newly formed floc particles through surface complexation and electrostatic attraction mechanisms. Once phosphorus has been effectively bound to these floc structures, the particle-bound phosphorus can be physically removed from the treated water through subsequent phase separation processes, including sedimentation in clarifiers, dissolved air flotation systems, or filtration through granular media or membrane barriers.

Operating on analogous physicochemical principles to those governing phosphorus removal, most heavy metals and trace metal species present in municipal and industrial wastewater also demonstrate a pronounced tendency to adsorb onto particle surfaces, precipitate under the elevated pH conditions often created by coagulant addition, or co-precipitate with the metal hydroxide flocs formed during the coagulation process. Consequently, these heavy metal contaminants will be removed from the aqueous phase through the same particle separation mechanisms that achieve phosphorus removal. The extent of metal removal generally correlates proportionally with the overall efficiency of particulate matter removal, with removal rates increasing as total suspended solids (TSS) reduction increases, though specific metal removal efficiencies can vary depending on solution chemistry, pH conditions, metal speciation, competing ligands, and the specific coagulant chemistry employed.

Our systematic investigation and analysis of treatment plant performance data yielded compelling evidence that most metal contaminants of regulatory concern, particularly copper (Cu), zinc (Zn), arsenic (As), cadmium (Cd), and lead (Pb), exhibited substantially lower effluent concentrations in

wastewater treatment facilities where chemical coagulation and precipitation treatment stages were present, when compared to facilities lacking this specific treatment step (Figure 4). The magnitude of concentration reduction varied by metal species but consistently demonstrated the effectiveness of chemical precipitation as a co-benefit for metal removal beyond its primary phosphorus removal function. These findings suggest that the implementation of chemical phosphorus removal technologies provides substantial ancillary benefits for controlling heavy metal discharge loads.

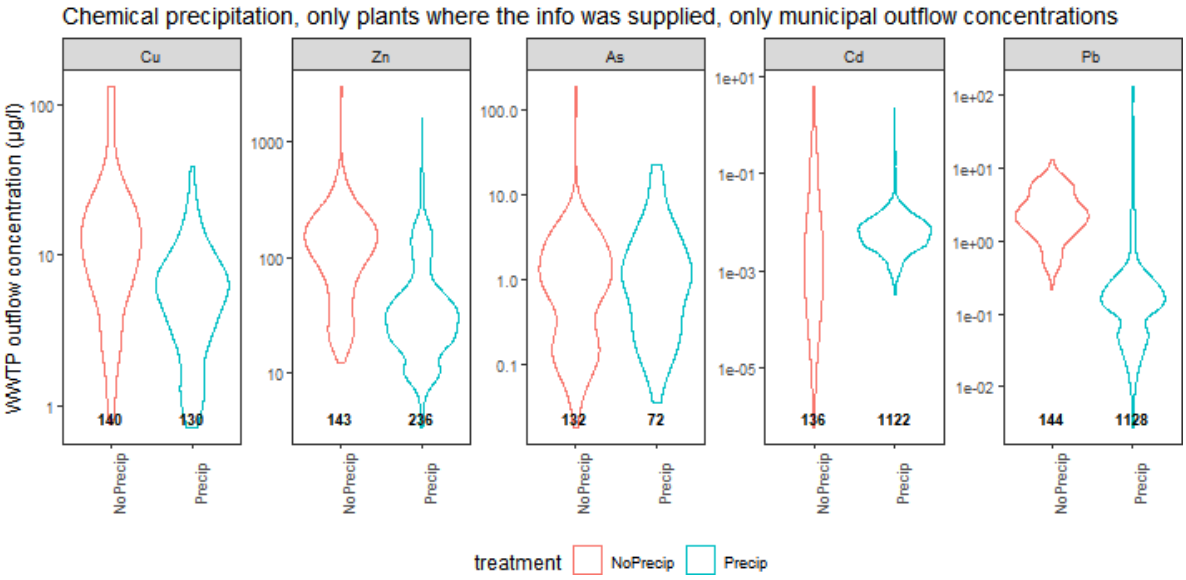

Figure 4: WWTP effluent concentrations by the presence / absence of a chemical precipitation step in the treatment technology. Number of measurements indicated.

Chrome and Nickel

For Cr and Ni, EF were derived based on plant size, because this subsetting indicated a relatively clear pattern. As a general tendency, larger plants tend to have higher industrial shares and thus higher concentration in heavy metals (Table 1).

Table 1 - Statistics and suggested EF of HM concentration by plant size (constructed capacity). Mean ± standard deviation (number of measurements > LOQ / number of total measurements)

WWTP capacity class	Statistics		Suggested EF	
	Cr	Ni	Cr	Ni
	µg/l			
PE < 2000	3.4 +/- 2.3 (12/16)	3.1 +/- 2.6 (24/28)	3	3
2000 <= PE < 5000	2.7 +/- 2.6 (17/29)	4.5 +/- 10.4 (41/51)	3	4
5000 <= PE < 10K	5.7 +/- 10.4 (14/19)	4.2 +/- 2.8 (92/96)	5	4
10K <= PE < 100K	4.9 +/- 7.8 (83/89)	5.4 +/- 6.0 (562/577)	5	5
PE >= 100K	4.6 +/- 8.1 (80/98)	6.1 +/- 4.0 (517/536)	5	6

2.5.1.2 PFAS

Two distinct approaches were employed to identify patterns in PFAS concentrations in wastewater treatment facilities: geographical regionalization and classification by plant size based on constructed capacity expressed in population equivalents. These complementary methodologies revealed different

but potentially interconnected factors influencing PFAS loads in wastewater streams across the study area.

For the geographical analysis, countries within the Danube River Basin were classified into three economic categories reflecting their development status and purchasing power. The "top-most" category comprised Germany, Austria, the Czech Republic, and Slovenia, representing the most economically developed nations in the basin. The "middle" category included Slovakia, Hungary, Croatia, and Romania, while the "down-most" category encompassed Serbia, Bulgaria, and Ukraine. Despite considerably more limited data availability from facilities in the "down-most" countries, the available evidence strongly indicated a clear gradient pattern: wastewater PFAS concentrations, measured in both influent and effluent streams, systematically decreased moving from the "top-most" through the "middle" to the "down-most" country categories. This spatial pattern suggests a plausible mechanistic explanation related to consumer behaviour and product availability—populations in economically more developed countries typically have greater purchasing power and access to high-technology consumer goods, particularly textiles, outdoor apparel, and accessories that contain elevated levels of PFAS compounds used for water-repellent, stain-resistant, or performance-enhancing properties (Figure 5).

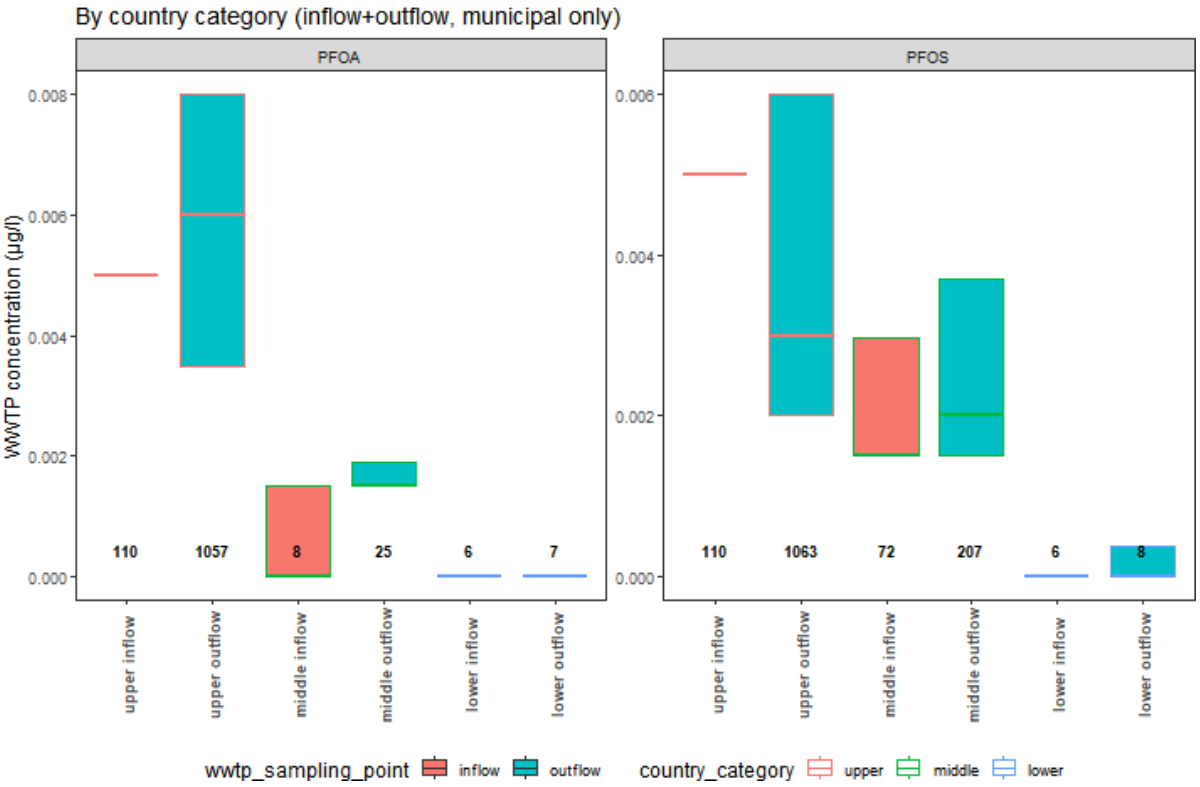

Figure 5: WWTP influent and effluent concentrations by country category. "upper" countries: DE, AT, CZ and SI; "middle" countries: SK, HU, HR and RO; "lower" countries: RS, BG and UA. Number of measurements indicated.

The second analytical approach involved stratifying treatment plants into five distinct size categories ranging from very small facilities with design capacities below 2,000 population equivalents to the largest installations exceeding 100,000 population equivalents, with effluent concentrations calculated separately for each capacity class. This size-based analysis revealed compound-specific patterns that differed notably between PFOA and PFOS. PFOA concentrations exhibited a U-shaped distribution, with elevated levels observed in both the smallest and largest plant categories, while intermediate and small-to-medium facilities showed comparatively lower concentrations. In contrast, PFOS

concentrations demonstrated a more straightforward monotonic relationship, increasing progressively with plant size category (Figure 6).

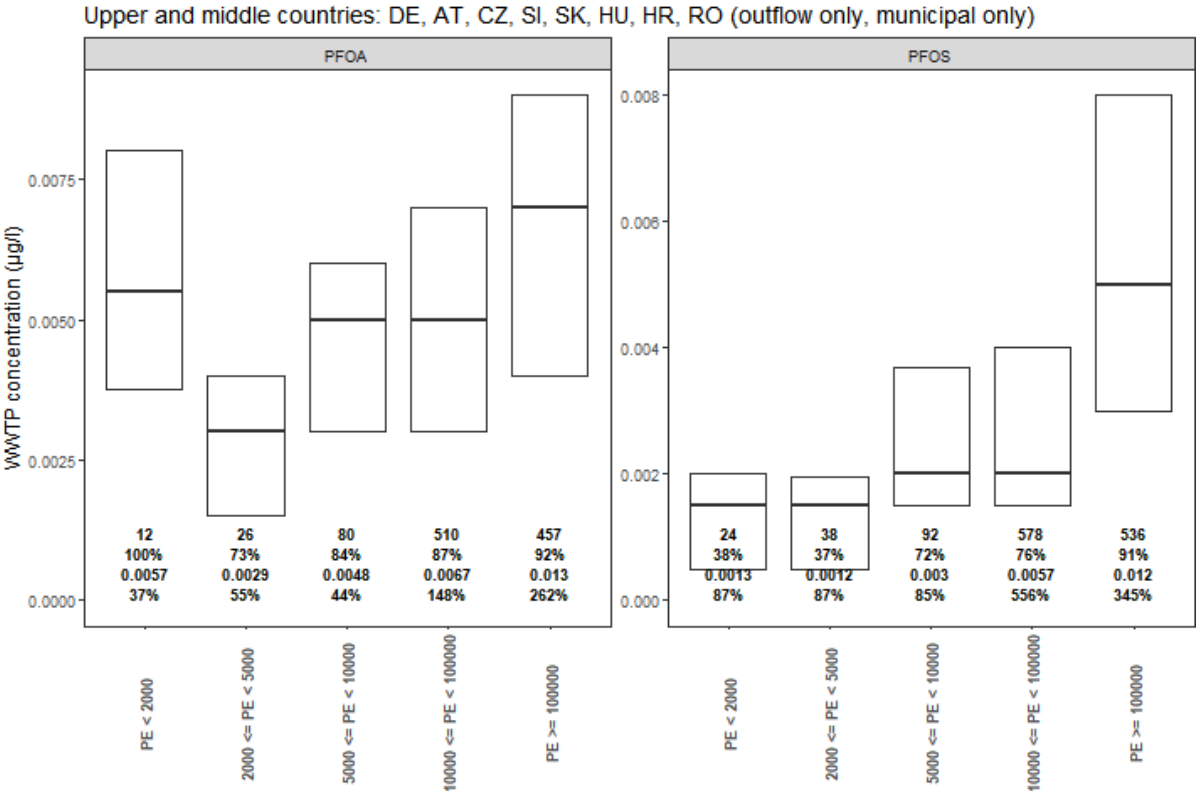

Figure 6: WWTP effluent PFAS concentrations by plant constructed capacity (p.e.).

These contrasting patterns likely reflect multiple underlying mechanisms operating simultaneously across the plant size spectrum. The elevated PFAS concentrations in the smallest facilities can be attributed to less sophisticated treatment technology and infrastructure limitations characteristic of small rural or suburban plants, which often achieve lower removal efficiencies for micropollutants. Additionally, these smallest plants may be disproportionately located in less economically developed regions or serve populations with different consumption patterns. Conversely, the elevated concentrations observed in the largest treatment facilities likely result from their receipt of substantial industrial and commercial wastewater contributions, including discharges from manufacturing facilities, textile finishing operations, metal plating industries, and firefighting foam contamination sources. Furthermore, the largest plants typically serve the most densely populated and economically developed urban centers where consumer purchasing patterns skew toward PFAS-containing products, creating an additive effect of both higher per-capita domestic loads and concentrated industrial inputs.

2.5.1.3 Pharmaceuticals

For pharmaceuticals, two further alternative methodological approaches were evaluated for developing emission factors for pharmaceutical compounds: (1) concentration-based emission factors expressed as mass per volume and (2) per-capita emission values expressed as mass per person per day. For each approach, values were calculated separately for individual countries participating in the monitoring program to identify and characterize potential regional differences. This dual analytical framework allows for flexibility in model implementation depending on available input data and the specific modelling objectives of different applications.

Analysis of concentration-based emission factors revealed substantial inter-country variability for the selected pharmaceutical tracers. Carbamazepine effluent concentrations ranged over two orders of magnitude across the study region, while diclofenac concentrations exhibited variability spanning one order of magnitude, indicating considerable heterogeneity in pharmaceutical loads among countries. Despite this overall variability, the majority of countries displayed effluent concentrations within a relatively consistent and narrower range of 0.1 to 0.5 µg L^{-1} for both pharmaceutical compounds, suggesting broadly comparable treatment efficiency and consumption patterns across much of the basin. However, notable exceptions to this pattern emerged in the dataset. Hungary exhibited substantially elevated concentrations for both carbamazepine and diclofenac compared to the regional baseline, potentially reflecting higher consumption rates, different prescribing practices, lower treatment removal efficiencies, or incomplete wastewater collection coverage. Conversely, Ukraine demonstrated substantially lower carbamazepine concentrations relative to other countries, which may result from lower pharmaceutical consumption rates associated with economic factors, differences in medical practice and prescription patterns, or alternatively could reflect limitations in analytical detection capabilities or sampling coverage (Figure 7).

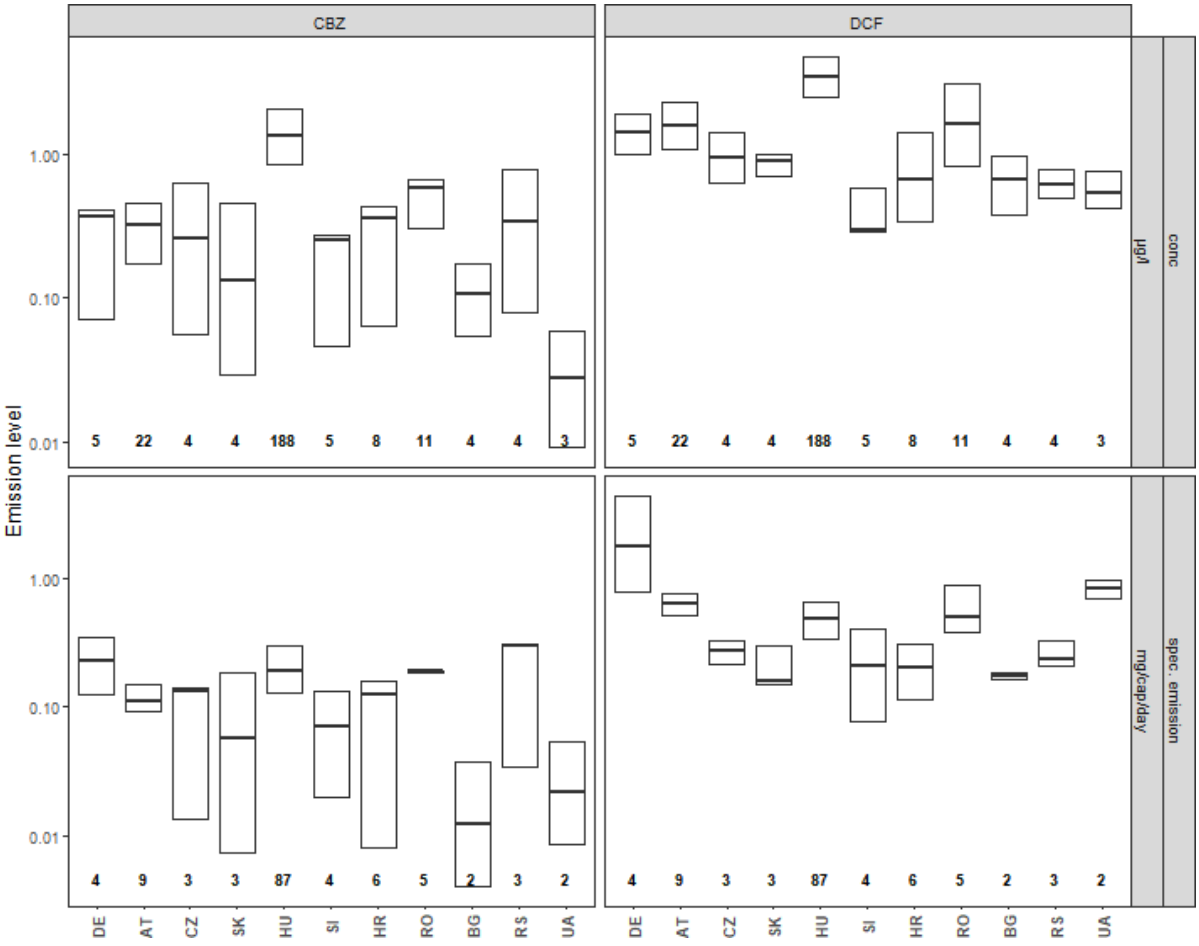

Figure 7: WWTP effluent concentrations and specific emissions by country. Number of measurements indicated.

The per-capita emission factor approach produced results within a generally narrower and more constrained range compared to concentration-based factors, suggesting that normalization by population served partially accounts for differences in wastewater flow rates and dilution effects among facilities and countries. Interestingly, the per-capita emission factors exhibited a discernible spatial gradient across the Danube River Basin, with a general tendency for values to decrease progressively from upper-Danube countries, particularly Germany and Austria, through mid-basin

nations toward lower-Danube countries including Bulgaria, Romania, Serbia, and Ukraine. This geographical pattern likely reflects the well-documented correlation between economic development and pharmaceutical consumption, with more affluent populations in Western European countries consuming higher per-capita quantities of medications, coupled with more comprehensive healthcare access and insurance coverage that facilitates pharmaceutical use. Both the concentration-based and per-capita emission factor approaches represent scientifically valid methodologies with distinct advantages for different modelling contexts, and consequently both can be effectively implemented within the modelling framework depending on data availability and the specific requirements of scenario analyses.

2.5.2 River concentration and load data for model validation

In the previous project Danube Hazard m³c, an inventory of hazardous substances concentrations was created, which is well described in Kittlaus et al. (2024). Most of this dataset was transferred into the Tethys transnational database. The spatially and temporally most extensive dataset covers surface water monitoring data in the DRB, including the transnational monitoring network (TNMN) of the Danube countries, managed by ICPDR. It operates since 1998, the first yearbook reporting heavy metal measurements was published in 2011. It contains mostly heavy metal measurements with monthly sampling frequency. Datasets for the modelling period and for 2021 were downloaded, covering all major tributaries monitored within the network.

The Joint Danube Survey (JDS) is an international scientific research expedition conducted along the Danube River and its major tributaries, organized by ICPDR every six years to collect comprehensive, harmonized data about the river's water quality, biodiversity, and pollution levels. JDS3, conducted in 2013 (i.e. earlier than the modelled period), placed significant focus on micropollutants with samples taken along the Danube River and at the confluence of main tributaries; and covered all project substances. JDS4, conducted in 2019 (i.e. within the modelled period), covered the Danube River, main tributaries, and even wastewater treatment plants, and provided thus a critical baseline for understanding micropollutant distribution. However, the lab analysis conducted by three independent laboratories led to substantially different results, making interpretation challenging.

The Tethys project's monitoring activity produced a spatially extensive dataset complementing the DH m^3c inventory. In non-EU countries (Montenegro, Bosnia and Herzegovina, Serbia, and Ukraine), sampling campaigns focused on river water (8–12 sites per country with six samples per site throughout the year. In EU countries (Austria, Slovakia, Slovenia, Bulgaria, Romania, Croatia, and Hungary), river sampling targeted outflow points of major tributaries, hotspots, and background locations, with each country collecting 4 samples per site under low-to-mid flow conditions across 2–3 sites.

The extended database’s data was used to validate the model results. Table 2 indicates the total number of available monitoring stations and the subset selected for validation.

Table 2 - Total number of available monitoring stations and the subset selected for validation were quantified for each substance.

Substance	Number of monitoring points selected for validation	Share of monitoring points with average concentrations > LOQ
Cr tot	239	75%
Ni tot	240	87%
Cu tot	236	95%
Zn tot	240	92%
As tot	237	86%
Cd tot	239	32%
Pb tot	240	76%
Cr diss	324	40%
Ni diss	326	80%
Cu diss	326	91%
Zn diss	324	86%
As diss	326	86%
Cd diss	318	31%
Pb diss	326	41%
PFOA	165	59%
PFOS	197	79%
CBZ	123	94%
DCF	138	91%

3 The Tethys pollution evaluation tool

The Tethys Pollution Evaluation Tool (PET) was developed to complement the transnational database with tools which support frequent tasks related to the database, e.g. supporting the quality check procedure by visualizing imported preliminary data to check their plausibility and compare them to other datasets in the database.

3.1 Functionality

The PET provides an integrated workflow for importing environmental data from the transnational database, preprocessing and analysing these data, generating exploratory visualisations, and producing automated quality-assurance reports. PET is designed to be modular and data-agnostic, enabling flexible use across multiple environmental compartments and data sources.

The functional scope of PET can be divided into three core domains:

1. Data Access and Import,

2. Data Visualisation and Exploration, and
3. Quality Assurance and Reporting.

Each domain is described in detail below. As the tool is implemented in the R programming environment[3], the tasks are conducted by calling R functions in an R working environment.

3.1.1 Data Access and Import

Database Connection and Project Initialisation

PET allows users to establish a connection to the transnational or national database through user-provided credentials (username, password, and optional role, database name, and host). Upon execution of the project initialisation routine (`load.project()`), PET configures the working environment and optionally connects to the database. Importantly, PET remains fully functional without an active database connection, if locally cached data are already available.

Import of Individual Tables

The function `pet_dt()` enables targeted extraction of table content from the compartment-specific database schemas of the Tethys databases. It supports:

- **Full table import** by specifying the schema and view name.
- **Column-restricted import**, achieved by providing a vector of required column names.
- **Row-level filtering** using SQL-like conditions supplied as character strings. PET supports multiple filtering statements (e.g., equality, membership via IN, combinations of conditions).

The output of `pet_dt()` is returned as an R object of the type data.table, facilitating efficient in-memory data handling.

Import of Merged Datasets

The function `pet_merge()` provides an automated mechanism for assembling complete data packages for specified data sources. It retrieves all relevant tables associated with one or more *identifier_datasource* entries and merges them into a unified dataset. Optional arguments (e.g., postfix for surface water and wastewater compartments) allow targeted selection of compartment-specific data layers. This procedure abstracts the complexity of the underlying schema structure, ensuring consistent dataset assembly across compartments such as soil, surface water, wastewater, and stormwater.

Subsetting and Outlier Detection

Because imported datasets are returned as *data.table* objects, subsetting operations can be performed efficiently using built-in syntax (e.g., filtering by determinants or sampling identifiers).

PET also provides an optional **outlier detection** routine via `pet_outlier_detection()`. This function flags observations lying outside a standard Tukey range defined by the 25th/75th percentiles ± 1.5 × IQR. A boolean column (outlier) is appended to the dataset to support subsequent filtering or reporting.

[3] https://www.r-project.org

Local Caching of Data

To enhance reproducibility and offline capability, PET offers a caching mechanism via `cache()`, which stores preprocessed datasets locally. Cached datasets are automatically reloaded in subsequent sessions, eliminating the need for repeated database imports.

3.1.2 Data Visualisation and Exploration

Overview of the Plotting Engine

PET includes a flexible plotting interface based on the `pet_plot()` function. This function dynamically constructs graphs using the R packages **ggplot2** and **plotly** using string-based arguments, allowing users to generate a wide variety of exploratory plots without extensive additional coding. All standard plot types (e.g., histograms, bar charts, violin plots, boxplots, scatter plots, line plots) are supported.

Plots are automatically saved to the PET/graphs directory when either a **title** or **loop identifier** is provided.

Core Plotting Features

The `pet_plot()` function supports:

- **Selection of axes variables** (x, y) and plot geometries (type, e.g., point, boxplot, histogram).
- **Aesthetic mappings** (color, fill) and freely definable additional aesthetics (free_in inside `aes()`, free_out outside).
- **Overlay of a secondary data layer**, defined through (y2, type2, color2, etc.).
- **Optional secondary y-axis** scaling (secaxis = TRUE), enabling dual-variable visualisations.
- **Facetting** across a categorical variable (facet_var).
- **Integration of custom ggplot statements** (free), allowing advanced modifications (e.g., adding reference lines).
- **Interactive or static output**, where interactivity can be disabled to enable compatibility with **patchwork** for composite plots.
- **Automatic log-scaling**, which may be overridden by the user.

Iterative Plot Generation

PET supports automated production of multiple plots using **for-loop–based iteration**. Users define:

1. the dataset to iterate over,
2. the categorical variable used for subsetting, and
3. the plot design applied to each subset.

Each iteration generates a standalone figure saved to disk, allowing efficient exploration of high-dimensional datasets (e.g., all determinants or all sampling sites).

3.1.3 Quality Assurance and Reporting

Dataset-Level Summaries

The tool provides automated reporting via the `pet_summary()` function. Given a merged dataset, PET produces an HTML report containing:

- descriptive statistics,
- completeness and metadata checks,
- optional filtering (e.g., summaries for outlier-only subsets).

Outputs are stored in the PET/reports directory. Users may customise the underlying RMarkdown template (lib/Summary.Rmd) to adapt summaries to specific QA requirements.

Data Source–Level QA Reports

PET extends dataset assembly and reporting through the `pet_QA()` function, which generates reports for selected data sources (naming: id_schema.html). This function internally performs a `pet_merge()` operation, ensuring that all relevant tables are incorporated prior to reporting.

The QA report provides a structured overview of:

- data source metadata,
- compartment-specific structure,
- variable availability,
- statistical summaries, and
- potential data quality issues.

The underlying template (lib/schema.Rmd) can be updated to accommodate evolving QA frameworks.

Summary of Functional Capabilities

In the current version, PET enables users to:

- **Access** databases of the type developed in Tethys with minimal configuration.
- **Import** entire tables or selectively filtered and merged datasets.
- **Preprocess** data through subsetting, outlier detection, and local caching.
- **Visualise** datasets through a flexible, parameter-driven plotting interface.
- **Automate** creation of exploratory and iterative graphics.
- **Generate** structured QA and summary reports for datasets and data sources.

Together, these functionalities establish PET as a modular and extensible platform for environmental data evaluation, designed to streamline workflows and improve transparency in pollution-related assessments.

3.2 Technical implementation

During the design of the PET, certain conditions were created to set a clear framework for the technical implementation to ensure a user friendly and expandable tool for further use:

The tool is implemented in the R programming language and thus build on opensource software, which is also freely available. The entire application should be compactly stored in a project template structure[4] to manage project environments, caching, and reproducibility. Different topics are in individual files in their designated folders. This should preserve a clear order, ensure a smooth workflow and enable easy distribution.

4

Codes should be clean, well-documented, robust, easy to expand, and structured into function chunks that perform one task each and can be used multiple times. In these functions the code should be wrapped in “tryCatch” functions which print unique error messages and should help significantly at error finding. This serves smooth understanding, expansion and error shooting of the code.

The code development platform Gitlab is applied to host the code and control the versioning, feature implementation and backup security in case of versioning issues or related compatibility errors.

This serves smooth understanding, expansion and error shooting of the code.

The information about every **A**pplication, **P**ackage, **E**xtension or specified **S**etting (APES) (such as ggplot2 prefixed settings) used in this tool’s version are documented in the README.txt file.

As this tool will use several APES at the same time, a strict version control and compatibility tests will be needed to ensure smooth development. To reduce the risk of errors due to incompatibility among APES, their quantity should be kept low if possible. Nevertheless, versioning will be mandatory.

The key functions of the tool for smooth connection with the data base, the data transfer between the applications, the automated data aggregation, the plotting and statistical functions as well as the automated report writing function are stored in scripts in the *PET/munge* directory and build the fundament of the PET. The scripts in *munge* have the task to fill the data files of the project accordingly as it is documented in the README.txt file.

The config file was then filled with all relevant information regarding the APES of the scripts, mostly stored in in the *PET/config/global.txt* file.

In order to ensure a smooth start for new users the README.txt file explains the set-up, the structure and best practice tips in ten steps leading to the three scripts long tutorial, that is stored in the *PET/scr* directory and explain all features the tool have.

The first script concerns itself with the import of data, the second script with the plotting of data and last with the creation of automated generated quality assessment reports of a data set.

After working through these 3 scripts the user should be able to easily import and merge data sets, visualize a high variety of data of distinct origin and speed up the quality assessment process.

3.3 Availability of the Tethys PET and next steps

As the tool builds on the structure of the Tethys database, it was developed after the main developments of the database and thus the first usable version became available. At present, the tool was tested by the project team and will be further improved based on the user feedback in the framework of the implementation of national databases in the parallel activity of the Tethys project that is currently still ongoing. The final version of the tool will be made publicly available on Gitlab under the following link:

https://gitlab.tuwien.ac.at/e226-1-working-group-river-basin-management/tethys

In parallel development of a more advanced version of the tool is ongoing, which will make use of the R Shiny package[5] which allows to program graphical user interfaces for the tool, which can then be used without any programming in a web browser.

4 Conclusions and outlook

The original HS concentration database developed within the Danube Hazard m^3c project was successfully and thoroughly revised, extended and optimized, and was thus transformed from an expert-tool into a much more operative management-tool. Among other aspects, the new version of the database has a significantly improved and more efficient structure, enhanced compatibility with other databases for EU reporting, extensive quality-checks and technical constraints as well as an accurate quality-control workflow. Moreover, a new tool to support the evaluation and processing of the data imported in the database was successfully developed.

The ICPDR is willing to adopt the transnational database (together with the emission model) as an operative tool to support periodic assessments of HS pollution related to transboundary river basin management. Furthermore, the ICPDR is willing to explore the possibility of integrating selected data into its water quality database to provide existing ICPDR databases with complementary information. The transnational database shall be hosted by an institution of the Danube countries, ideally strongly linked with the transnational emission model. Future database maintenance work and database updates shall be linked to the next model application and carried out by the hosting institute and the modellers, in close coordination with the ICPDR.

[5] https://shiny.posit.co/

Annex

Standard operating procedure to set up transnational HS database

Purpose of this document

This document should help administrators to set up the Tethys concentration database system at their server. Basic database and server administration knowledge is required, everything specific to this data base is explained in this document.

Prerequisites

What is required to install the data base.

Server hardware

If a hardware server is set up or a virtual server is created is depending on the available IT infrastructure, both is working well, if implemented correctly.

The requirements for a hardware machine are

- processor: 6 cores or more, 1.2 GHz or more
- working memory: 16 GB or more
- free hard disk space: 100 GB or more, of which min 50 GB on fast partition (SSD).

On a hardware Server for data security and speed considerations the use of a hard disk RAID is recommended.

Software

The database is implemented in the PostgreSQL database management system (DBMS). PostgreSQL is open source and can be retrieved from: https://www.postgresql.org

The database was developed with PostgreSQL version 15.2, this and newer versions should be suitable to run the database.

During database installation for our purpose default settings are fine. For the write ahead logs (WAL) the fast partition should be chosen.

As the database makes use of the PostGIS extension (https://postgis.net), this should be installed after installation of PostgreSQL. On Windows operating systems the PostgreSQL stack builder application can be used to install the extension (Figure 8).

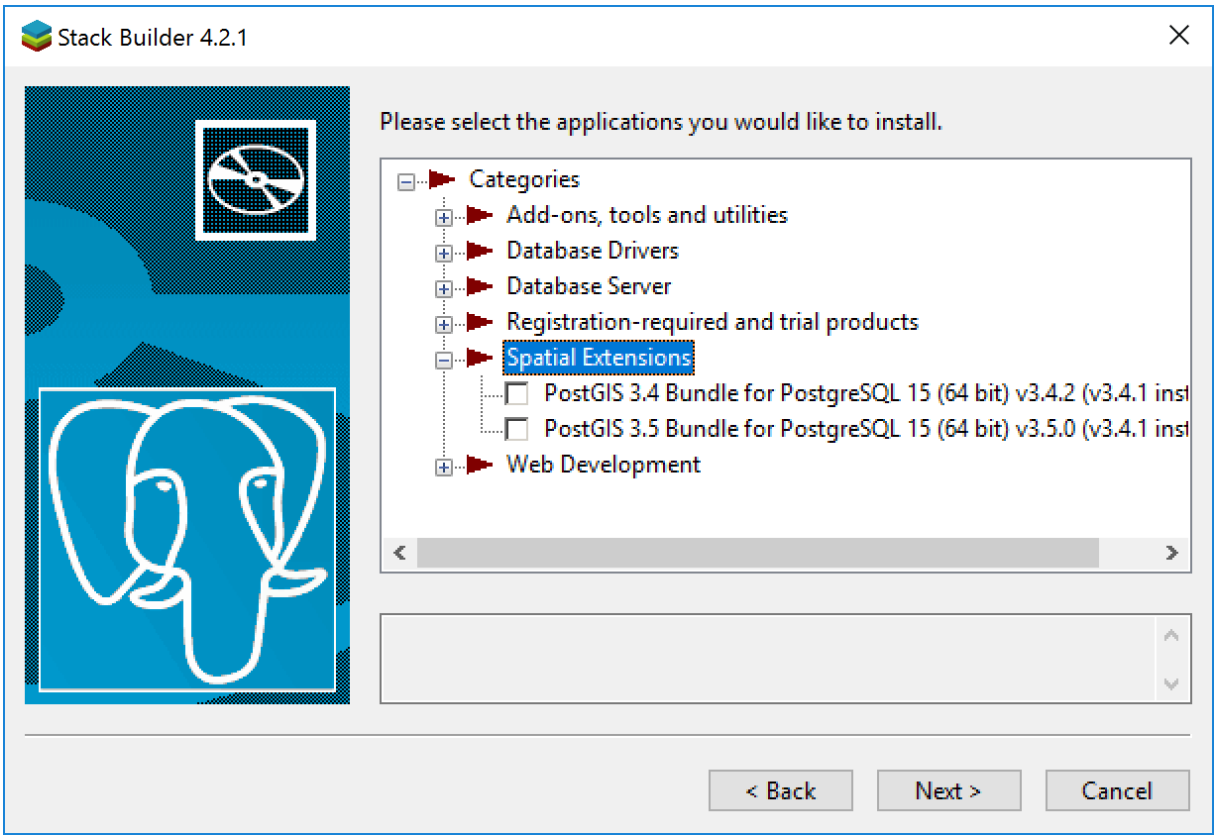

Figure 8: Install PostGIS extension with the application stack builder.

Setup of the database

The setup SQL scripts are available in one folder numbered in the sequential order they need to be run.

The administrator needs to connect to the PostgreSQL DBMS using any available database, if Default setup was chosen a database named “postgres” is available and the default superuser is named “postgres” as well.

As a superuser the administrator needs to run script 00_Create_Roles_and_DB.sql which creates the necessary roles and the database. Make sure to insert the name you want to give for your database.

For the script 01_Create_schemata_add_extension.sql the Administrator needs to connect to the new database still using the superuser role.

Starting from script 02_ERmaster_exported_DDL_script.sql the administrator can use any user which has been granted the tethys_admin role.

Once all scripts have been successfully executed, the database including the controlled vocabularies is ready for usage.

Documentation

The documentation and a FAQ are available under

https://gitlab.tuwien.ac.at/e226-1-working-group-river-basin-management/tethys/tethys-db-schema/

The documentation includes a detailed description and instructions of the following topics:

- Concept: brief introduction of the conceptual structure of the database
- Included data: list and short description of the environmental compartments whose data and metadata can be imported in the database
- Structure: List, description and explanations of the schemata covering the different specific environmental compartments, the available views fulfilling different purposes
- User rights management: list and explanation of the role and functions of each specific user type
- Data quality assurance: description of the concept and detailed workflow of the implemented procedure for data quality assurance
- Order of data import: detailed description of the order to be followed to import new data, to support an efficient and workflow given the complexity of the database structure
- Technical aspects not important for basic users: advanced information for IT and Database experts

Standard operating procedure to backup transnational HS database

Purpose of this document

This document should assist administrators in backing up the Tethys concentration database system on their server. Basic database and server administration knowledge is required; everything specific to this data base is explained in this document.

Prerequisites

What is required to backup the database:

User Access:

Ensure that users executing the backup have the necessary permissions to access *pgAdmin* and perform backups on the database.

Software

The *pgAdmin* tool must be installed on the system where the procedure will be executed. It can be downloaded from https://www.pgadmin.org/download/

Hardware

Determine and have access to a secure location for storing backup files. This could be local or network storage.

Backup Procedure

Database Backup

1. Launch the *pgAdmin* application on your computer
2. Connect to the database server
   In the browser panel on the left, expand the server section by clicking on it.
   Right-click on the database you want to back up (Figure 9).
3. Start the backup process
   From the context menu, select “Backup....” (Figure 9)
4. Configure the Backup Options
   In the Backup Database dialog “General”, fill in the following fields:
   Filename: Specify the path and name for the backup file
   (e.g., D:\backups\database_ABC_YYYYMMDD.backup).
   Format: Choose the backup format (select Custom).
   Encoding: Select the encoding (usually UTF8).
   In the Backup Database dialog “Data Options”, two options are available:
   For only data backup: Under „Type of objects“, ensure that „Only data“ and „Blobs“ are checked.
   For full backup: Under „Type of objects“, ensure that only „Blobs“ is checked
5. Initiate the Backup
   Click the Backup button at the bottom of the dialog.
   You will see a progress dialog indicating the status of the backup process.
6. Confirm Completion
   Once the backup completes, you’ll see a confirmation message.
   Check the specified folder to ensure that your backup file has been created.

For a detailed guided process through the *pgAdmin* interface, please refer to the official documentation: https://www.pgadmin.org/docs/pgadmin4/development/backup_dialog.html

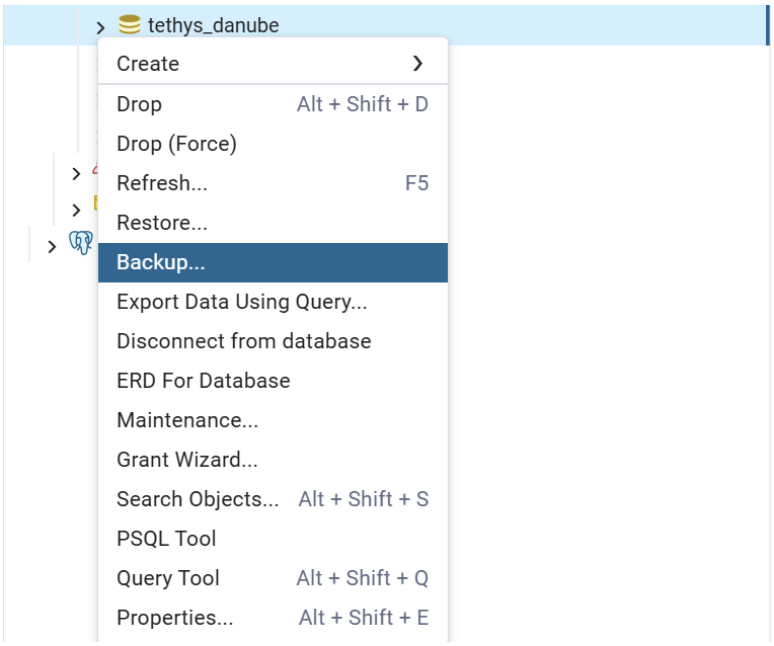

Figure 9: Right-click on the database you want to back up