Computers & Geosciences 190 (2024) 105640

Contents lists available at ScienceDirect

Computers and Geosciences

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

Research paper

PHREESQL: A toolkit to efficiently compute and store geochemical speciation calculation

Marino Vetuschi Zuccolini [a,b,*,1], Daniela Cabiddu [b,1], Simone Pittaluga [b]

[a] *Lab. of Geochemistry - DISTAV - University of Genova, Corso Europa, 26, 16132, Genova, Italy*
[b] *CNR-IMATI, Via De Marini, 6, 16149 Genova, Italy*

ARTICLE INFO

Keywords:
Hydrogeochemistry
Speciation computation
Relational database
Big data storage and query

ABSTRACT

Both abiotic and biotic natural spheres benefit from the high reactivity of natural waters, which are ubiquitous on planet Earth. The use of speciation-solubility codes like *PHREEQC* can provide a deeper understanding of aqueous equilibria and water-rock interactions. A significant number of newly derived variables are produced by these computations, which may significantly increase compared to input data. It is crucial to process vast amounts of data efficiently, particularly when dealing with datasets that contain thousands of water analyses.

To tackle this problem, we present *PHREESQL*, a software package designed to efficiently store and manage extensive data generated by geochemical speciation computations performed by *PHREEQC*. High efficiency in data extraction and filtering of the entire output from a single run can be achieved through a well-designed relational SQL database structure. The *PHREESQL* can be used as a stand-alone package or embedded in third-party applications. Thanks to the SQL structure, it is possible to create links with unstructured meshes by developers and experts in reaction-transport problems. Real-time data management from multiparameter devices in field and laboratory settings is made possible and efficient by parallel computation options and software integration. The toolkit encompasses both a C++ library and a command-line interface, facilitating its use by geochemists with limited programming skills.

1. Introduction

In recent decades, geochemical speciation simulations have become an essential tool to deepen knowledge of aqueous systems. Several approaches and models have been proposed (Steefel et al., 2014; Leal et al., 2017) implementing thermodynamics databases built to be compliant with specific aqueous activity coefficient models (i.e. Debye-Hückel, B-dot, Pitzer or Brønsted-Guggenheim-Scatchard (SIT) (Blanc et al., 2012; Giffaut et al., 2014; Lothenbach et al., 2019; Lu et al., 2022)). Among them, numerical methods based on mass-action law implemented in software like *PHREEQC* (Parkhurst and Appelo, 1999, 2013) generate results whose output includes a large number of new derived variables.

The number of variables stored in the output (i.e., saturation indices for phases, activity, and activity coefficients for species) may increase by orders of magnitude compared to the ones included in the input data (i.e., molalities, and physical parameters). Such an amount of information, when water analyses are more than a few, could benefit from post-processing tools capable of making effective use of all the performed simulations.

Bio-geochemists can explore PHREEQC solubility-speciation capabilities using different strategies depending on their personal skills. Inputs for computations are produced by using graphical user interfaces (i.e. Phreeqci (Charlton et al., 1997)) either by creating files with a full syntax compliance or typing input files and running them by a command line. Since the two options require user text manipulation, they are prone to errors and are time consuming. A "one-to-one" relation between input and output is the commonest game plan to compute a few number of speciation calculations. Dealing with a large number of water analyses can be challenging, and PHREEQC overcomes this limitation, changing the paradigm (from "one-to-one" to "many-to-one") and providing the possibility to run a serial computation of more than one analysis and generating a single output ASCII file. The size of such an output can be very large, which lead to inefficient and time consuming data extraction. Although PHREEQC enables the extraction of a subset of variables from the output file, such a strategy is limited by the impossibility of extracting data other than those already available without running a new simulation. In the presence of hundreds or thousands of water analyses, the set of calculations

* Corresponding author at: Lab. of Geochemistry - DISTAV - University of Genova, Corso Europa, 26, 16132, Genova, Italy.
E-mail address: marino.zuccolini@unige.it (M. Vetuschi Zuccolini).
[1] Joint first authors.

https://doi.org/10.1016/j.cageo.2024.105640
Received 6 April 2023; Received in revised form 8 February 2024; Accepted 29 May 2024
Available online 6 June 2024
0098-3004/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

cannot be completed without a significant effort, including the waste of time and the high error rate due to the strict syntax, induced by keyboard typing. It is evident that making the process automatic by the development of dedicated software solutions would be advantageous, but not all users have required programming skills to achieve the goal.

Additionally, PHREEQC has capabilities for one-dimensional (1D) transport calculations and can simulate 2-3D domains by integrating external transport codes such as HT3D (McGrattan et al., 2020) and *PHAST* (Parkhurst et al., 2010). Results of such simulations determine the spatial distribution and the temporal variations of all parameters of interest. Again, results (saved in binary form, i.e. according to HDF5 standard (The HDF Group, 2000–2010)) are strictly limited to what is declared in the input file and then additional parameters of interest should be extracted by recalculating. The information density of geochemistry over space and time should be supported by a data storage solution extending the data mining capabilitys.

Our contribution in this article is to present *PHREESQL*, a toolkit for efficiently storing, retrieving, and querying water speciation calculations from *PHREEQC*. *PHREESQL* includes a C++ library, namely *PHREESQLib*, to be exploited by developers while coding their own tools to perform *PHREEQC* computation. Also, a command-line interface, namely *PHREESQLexe*, is provided to enable and simplify the exploitation of *PHREESQL* by users who need to use the tool with low programming expertise.

The paper is organized as follows. Section 2 clarifies the motivations of our work, while Section 3 provides reminders about geochemical speciation simulation in *PHREEQC* framework. Technicalities about the proposed toolkit are described in Section 4, while implementation aspects are described in Section 5. As a matter of an example and usage demonstration, a case study is in Section 6 and conclusions are finally discussed in Section 7.

2. The needs for *PHREESQL*

Nowadays, large datasets related to water geochemistry are integrated into online databases and are widely distributed. The vast majority of the databases are of national interest and are intended to report compositional data about groundwater and surface water quality and budget evaluation. Some examples of online hydrogeochemistry databases can be found in literature (USGS, 2016; UFAM, 2023; Federal Ministry Republic of Austria, 2023; Geological Survey of Ireland, 2023; EPA-IE, 2023; EPA-USA, 2023; WRIS-India, 2023). The scientific community can now process and integrate data from independent sources on a level never seen before, thanks to the abundance of large and freely distributed datasets.

Water geochemistry datasets, both surface and groundwater, come primarily from two main pipelines:

- by merging unsupervised databases, often with little in-house coherence, into a single supervised database;
- by populating a database while sampling environmental variables through a (possible multi-sensor) data source.

The first pipeline is the most common procedure where national or regional environmental agencies are responsible for collecting a variety of digital datasets or data reports or papers, turning them into monolithic digital files (as in ASCII, SQL or proprietary data formats). It usually takes a long time to converge to a final form, requiring a huge effort to standardize different databases into a single one, a supervised quality and control analyses, and an error-prone analysis of data consistency. Due to the process' complexity, the delayed-in-time activities done by various analysts may introduce some errors into the output database. Such inconsistencies could be attributed to typing errors or inconsistent units of measurement.

In contrast, the second pipeline enables the possibility to collect datasets through an embedded hardware-software infrastructure. Data collection is intended to be performed in accordance with a customized protocol (Caccia et al., 2019), and it enables the possibility to create huge, dense, heterogeneous digital datasets as the same time as the data acquisition takes place. Unfortunately, the real-time dataset storage approach is not standardized, and it is necessary to design it properly to guarantee both storage and query efficiency.

Although consistency and completeness of water analyses in a dataset are unavoidable features to approach the comprehension of an aquifer status, the description of equilibria and water-rock interactions needs an extra-effort. If distributed datasets are updated periodically but at a very low frequency, a continuous high-frequency data streaming can be generated during experiments, involving both sensors, digital acquisition and communication systems (Pötter et al., 2021). Examining a dataset on one hand and utilizing *PHREEQC* as the solver for a single-point speciation calculation on the other, there should be a link capable of pre-processing, storing, and visualizing results promptly.

PHREESQL is designed to support both the storage and query of water geochemistry datasets processed by *PHREEQC*. It helps experts persistently store, update, and query the results of the aqueous speciation simulation through the implementation of a relational database management system (RDBMS). The structure of the database has been properly designed to guarantee efficient data storage, reduce the amount of disk space and enhance query performances.

3. Background

In the framework of geochemical speciation simulations based on the mass-action law, *PHREEQC* (Parkhurst and Appelo, 1999, 2013) is one of the most commonly used packages. It allows numerically approximate solutions to a wide range of problems in aqueous geochemistry. The Open Source nature of *PHREEQC* has kicked off many projects, including but not limited to *iPHREEQC* (Charlton and Parkhurst, 2011), *PhreePlot* (Kinniburgh and Cooper, 2011), *PHAST* (Parkhurst et al., 2010)), and various implementations or interfaces (PhreeqPy (Müller and D.L. Parkhurst, 2011), R (Charlton et al., 2022), PEST (Mosai et al., 2022), Rich-PHREEQ (Wissmeier and Barry, 2010), COMSOL (Guo et al., 2018), OPENFOAM (Pavuluri et al., 2022), PHT3D (Appelo and Rolle, 2010), UTCHEM (Korrani et al., 2015).

Specifically, *PHREEQC* has been wrapped into a C++ library, namely *iPHREEQC*, which can be included in programs written in C++ (Charlton and Parkhurst, 2011). The modularity of *iPHREEQC* allows easy implementation of parallel processing for computationally intensive geochemical simulations. In the following sections, the term *PHREEQC* refers to its C++ implementation in *iPHREEQC*.

PHREEQC is designed to be flexible enough to implement various strategies for processing a large number of water compositions. It serializes results into ASCII files, including a large number of newly derived variables. The total size may increases by orders of magnitude from the input data. Experts have to extract the relevant information through each single output manually or through dedicated programming solutions. Although *PHREEQC* allows customized file format outputs by reducing variables relevant to the case-study, extracting information can be time-consuming due to numerous generated files (one for each simulated water analysis). In addition, such a manual post-processing may be prohibitive or hopeless if data is collected *in-situ* by sensors and the speciation calculation needs to be performed in *real-time* in order to adaptively modify a geochemical sampling schema as in (Berretta et al., 2018).

To the best of our knowledge, data extraction can only be significantly accelerated by implementing ad-hoc solutions. Such an implementation requires advanced programming skills, and is often difficult to be developed by geochemists without a proper programming background.

PHREESQL aims to be a robust code, especially given the high information density stored in the available datasets. In fact, it is designed to provide the capacity to process a huge amount of data ensuring both persistent storage and query efficiency. Also, *PHREESQL* is easy to use by either geochemical or programming experts, since it is provided as a combination of a code library and a standalone application.

4. Toolkit description

PHREESQL is an easy-to-use framework supporting the efficient storage of *PHREEQC* speciation calculations.

The production of a relational database is the core of the toolkit, making it possible to permanently store, in a structured framework, a very large number of geochemical calculations, compacting all data into a SQL binary form.

To guarantee usability for both developers and users with low programming expertise, *PHREESQL* includes both a C++ library and a command-line interface, namely *PHREESQLib* and *PHREESQLexe* respectively.

Thus, *PHREESQL* exhibits the following capabilities:

- *run PHREEQC speciation calculation*;
- *build* a SQL database from scratch and *update* an existing database with additional data;
- *export* both original *PHREEQC* inputs and outputs from an existing database for further processing;
- *process* coordinates of geo-referenced data into an existing database *to create* a new table or to *create* a new database framed into a desired specific coordinate reference system (CRS).

The following sections provide details on *PHREESQL*'s database structure (Section 4.1), exposed main functionalities (Section 4.2) and the supported input file format (Section 4.3). Note that these aspects are common to both *PHREESQLib* and *PHREESQLexe*, thus the term *PHREESQL* refers to both of them.

4.1. Database structure

A *PHREESQL* database is a SQL relational database consisting of a collection of tables storing structured data coming from geochemical speciation simulations run by *PHREEQC*. Each table includes a collection of rows, also known as records or tuples, and columns, also referred to as attributes.

More precisely, a *PHREESQL* database is structured as a collection of 7 tables reflecting the syntax of *PHREEQC* inputs and the sequential data stream of *PHREEQC* outputs (see Fig. 1 and Fig. 7). The primary table, namely `METADATA`, stores information about each analysis (location, sampling time and *PHREEQC* options) in each row. In this table, the *primary_key*, namely "ID", serves as the unique analysis identifier, essential for joining the primary table with all the others. Besides the primary table, a table named `SOLUTION_INPUT` stores, for each input analysis, data from the *PHREEQC* input file. Five additional tables are responsible for storing data from the *PHREEQC* output and reflect the original structure within the output ASCII file:

- the table `SOLUTION_COMPOSITION` stores the chemical composition in molalities for each input analysis;
- the table `DESCRIPTION_OF_SOLUTION` provides a basic summary of solution's chemical and physical properties;
- the table `DISTRIBUTION_OF_ALKALINITY` stores the alkalinity contribution of each species in solution;
- the table `DISTRIBUTION_OF_SPECIES` contains the main results from the speciation computation, listing all the species in solution;
- the table `SATURATION_INDICES` stores information on solution equilibrium relative to mineral phases reported in the thermodynamic database.

4.2. Main functionalities

PHREESQL mainly provides four functions ranging from data preprocessing to data post-processing.

4.2.1. Running speciation calculation.

To build a *PHREESQL* database, both inputs *.pqi* and outputs *.pqo* of *PHREEQC* are mandatory. If the output is unavailable, *PHREESQL* provides the option to generate it through a direct call to *PHREEQC*. This procedure can process either a single input file or recursively all speciation calculations stored in a single folder. For each input file, a corresponding output file is created and saved in the provided output folder.

Thanks to parallel programming techniques (Chandra et al., 2001), this procedure is much more efficient than running *PHREEQC* iteratively in a loop within a shell script (see Section 6.2).

4.2.2. Database creation and/or update.

The main functional performance of *PHREESQL* is to insert metadata and *PHREEQC* input/output pairs into a SQL relational database. If the database does not exist, it is created as the initial step of the procedure. For each input analysis, the internal procedure checks if it already exists in the database, verifying if a record with the same *.pqi* filename, date and *PHREEQC* database exists as trigger. If the answer is affirmative, the speciation computation is skipped and the database remains unaltered. Conversely, if the answer is negative, both analysis metadata and the *PHREEQC* input/output pair are processed, and the data are inserted into the database, referencing the same chemical composition.

Note that the same *PHREEQC* input data can be inserted into the SQL database multiple times if multiple speciation runs have been calculated with different *PHREEQC* thermodynamic databases.

4.2.3. Data export

PHREESQL enables the export of data to ASCII files for further processing. There are several options for exporting one or more analyses. For example, *PHREESQL* provides the possibility to export both *PHREEQC* input and output files. This export adheres to the *PHREEQC* data format with a sort of file structure back-transformation, thus avoiding the storage of the original *.pqi* file.

4.2.4. Coordinate conversion/transformation.

As potential geo-referenced data, geochemical analyses must be supported by a flexible tool for coordinate conversion and transformation. For each analysis, *PHREESQL* reads coordinates and the European Petroleoum Survey Group (EPSG) (Managed by the International Association of Oil and Gas Producers IOGP, 2023) code from the input metadata file and stores it in the `METADATA` table (see Section 4.3). This procedure allows the storage of reference systems of geographical or projected coordinates and enables the coexistence of sampled analyses in different coordinate reference system.

If there is a requirement to store spatial information according to a specific EPSG code, *PHREESQL* can perform on-demand conversion. This is achieved through the specification of a new EPSG target in the command line together with the specification of how the result of such a conversion must be stored. Specifically, *PHREESQL* allows the possibility to export the result of the transformation into three different ways, that is by creating one of the following outputs:

- a *.csv* file encoding a copy of the `METADATA` table where the values in columns `COORD_X`, `COORD_Y` and `EPSG` are updated according to the desired reference system;
- a new *PHREESQL* database which is a copy of the original one, but having the `METADATA` table updated according to the desired reference system;
- a new table in the original database, which is a copy of the original `METADATA` table updated according to the desired reference system.

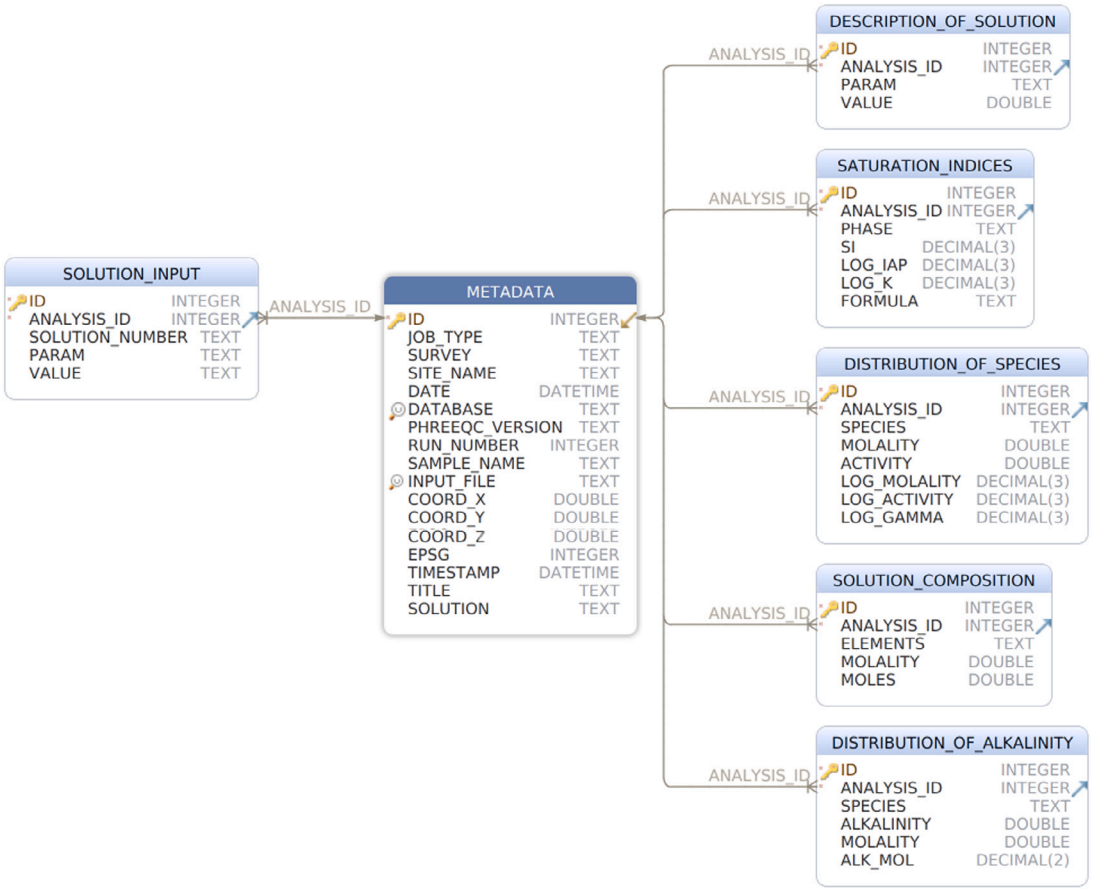

Fig. 1. *PHREESQL* database structure, consisting of a collection of 7 tables reflecting the syntax of PHREEQC inputs and the sequential data stream of PHREEQC outputs. The primary table, namely `METADATA` (center), stores information about each analysis (location, sampling time and PHREEQC options) in each row. The primary_key, namely "ID", serves as the unique analysis identifier to join the primary table with all the others. Besides the primary table, a table named `SOLUTION_INPUT` (left) stores, for each input analysis data from the PHREEQC input file. Five additional tables are responsible for storing data from the PHREEQC output and reflect the original structure within the output ASCII file (right). The complete state of a database is usually contained in a single binary file on disk, according to SQlite3 specifications.

4.3. Input data

Creating and/or updating a SQL database by *PHREESQL* requires at least one geochemical analysis. A collection of analyses is supported, as long as they are available in a single folder.

Every chemical analysis must consist of a pair of files:

- a *.pqi* file encoding the *PHREEQC* input according to its syntax. The structure of the *PHREEQC* file must include a single speciation calculation in a SOLUTION-END data block. Optionally, the PRINT clause can be considered, allowing the computation and decomposition of the total alkalinity by each species;
- a *.met* file, which is an ASCII file with additional information (other then chemical data, see Appendix A). The metadata file increases the completeness of the analysis description, as shown in Listing 1. Metadata present in the SQL database enables a more extensive integration with external software packages, such as plotting a graph or the implemention into GIS spatial databases.

Listing 1: An example of the Metadata File Format (*.met*)

```
JOB TYPE: Test4PHREESQL
SURVEY: Limerick
SITE: Doon
DATE: 2014/11/03
DATABASE: llnl
PHREEQC_VERSION: 3.6.1-15000
RUN_NUMBER: 1
SAMPLE_NAME: IE_SH_G_0213_3600_0008
INPUT_FILE: EPA_09894.pqi
COORD_X: 181429
COORD_Y: 150645
COORD_Z: 10
EPSG: 29902
TIMESTAMP: 1679786302
```

As aforementioned, *PHREESQL* also requires the output of the *PHREEQC* speciation calculation, provided as an ASCII *.pqo* file. If it is already available, the user can provide it as a third optional parameter. Otherwise, *PHREESQL* is able of generating them.

A dataset consisting of several analyses can be supplied as input. In this case, the entire dataset must be available as a set of folders, namely *input* and *metadata*, each of which includes all the *.pqi* and *.met* files respectively. Intuitively, a third folder can be provided, namely *output*, including previously generated *.pqo* files related to *.pqi* files in the input folder. When a dataset includes several water analyses, *PHREESQL* assumes that the same basename is used to name the pair (or triplet) of files related to the same single analysis.

5. How to use *PHREESQL*

PHREESQL is released as a standalone open-source toolkit, available via Github (https://github.com/DanielaCabiddu/PHREESQL). It includes both a C++ library and a command line interface, making it usable by both developers and in practical applications. The following sections provide implementation details and technicalities necessary

[for using] either the library (Section 5.1) or the command line interface (Section 5.2). The source code is documented with Doxygen, providing both HTML and LaTeX documentation for additional technical details. This documentation supports users in searching for specific features. *PHREESQLexe* provides a help function to display a concise description of the program and each available option, covering syntax and synopsis. Examples illustrating how to use the toolkit are provided together with the toolkit itself.

5.1. PHREESQL C++ library

PHREESQLib is a header-only C++ library, making it easy to use by simply including the headers into the user's project. No compilation, packaging and installation is required to use it, although some package dependencies must be installed for both compilation and execution.

To run, *PHREESQLib* relies on external libraries to implement some functionalities. Specifically, it uses SQlite3 (Hipp, 2020; Owens, 2006) for generating and managing the SQL database. Furthermore, *PHREESQLib* depends on *iPHREEQC* for running geochemical speciation simulations and on *PROJ* library (PROJ contributors, 2022) to transform or convert coordinates. These dependencies are mandatory, and must be installed on the local machine where *PHREESQLib* will be used.

Finally, *PHREESQLib* takes advantage of the potentials of OpenMP (Chandra et al., 2001) to parallelize and optimize the execution of some procedures, such as the run of geochemical speciation computations. While this dependency is not mandatory, it is strongly recommended to guarantee optimal computation performance.

5.2. PHREESQL command line interface

PHREESQLexe serves as the command-line interface of *PHREESQLib*, designed to support users with low programming expertise. *PHREESQLexe* is easy to use, as users can execute specific actions by providing the desired functionality as an argument, choosing from the available ones:

- `--run_phreeqc` to perform speciation calculation;
- `--fill_db` to either create a new database or update an existing one;
- `--export_input`, `--export_output`, `--export_metadata` to export inputs, outputs, and metadata, respectively, from the database to text files;
- `--epsg_convert` to convert geo-referenced data to a specific coordinate system.

Each action requires additional arguments to work correctly, such as the *PHREESQL* database reference and the path of inputs and outputs. Table 2 in Appendix B provides the comprehensive set of options and arguments for each action. Additionally, users can access the complete set of option by providing the `--help` argument to *PHREESQLexe*.

5.3. Data pre-processing

PHREESQL assumes the input data consists of one or more geochemical analyses, each consisting of a *.pqi* and a *.met* file, as detailed in Section 4.3. In practical scenarios, meeting this assumption might pose challenges, but it is feasible that a *.csv* file is available, encoding the entire set of analyses.

To support the user in generating PHREEQC inputs and metadata files starting from a *.csv*, the online distribution includes a set of bash scripts. The practical application of these scripts is illustrated in our case study, detailed in Section 6. Technical documentation for these scripts is provided in the online distribution for reference.

6. Case study

The application below has been designed to show how *PHREESQL* can be employed. Through some bash scripts, the complete pipeline for pre-processing input files, computing solution speciation, creating various SQL databases, and extracting data is demonstrated, starting from an input *.csv* file. The examples are based on a large dataset of publicly accessible geochemical data, covering the generation and post-processing of *PHREEQC* output data (see Section 6.1).

The entire procedure, implemented in scripts, is available in the distribution and can be evaluated at each step (see Section 6.2). Finally, we provide some examples of how to exploit PHREESQL databases to perform data extraction, including coordinate conversion for variographic study and mapping of a geochemical variable as output of *PHREEQC* computations (see Section 6.3), with the final aim to demonstrate the possibility to plot extracted data by the integration of *PHREESQL* with external software.

6.1. EPA Ireland groundwaters dataset

In order to involve citizens and stakeholders in government practices, environmental agencies share environmental data that has been hidden until now. To convert old data, collected over decades and stored in paper format, to digital format, it is required an enormous amount of effort and careful practices.

The Ireland EPA database[2] is a valuable effort to distribute the knowledge about groundwater quality throughout the entire country. The dataset is distributed with a proprietary digital file format (Excel©by Microsoft) formatted for easy use in capturing a snapshot of Ireland's water geochemistry (McGrory et al., 2018, 2021b).

The dataset contains a total of 295 sampling stations in Ireland's aquifers, which have been sampled for a total of 14690 analyses over a 30-year period until 2020. A total of 303 variables are present in the original dataset, which provide descriptive information, spatial coordinates, dissolved components or species, and chlorinated organic compounds. The file used as an input of our experimental pipeline is a post-processed version of the original Ireland EPA dataset, reduced to 45 columns ignoring the extremely sparse organic compound records. Specifically, the following changes have been applied:

- the creation of a new column which unambiguously identify a single analysis, to be included in the *PHREEQC* file name (i.e. the first analysis is labeled with an ID as 00001; 5 digit to consider up to 99999 samples). The existence of such an ID is mandatory for our procedure;
- the identification of data below the detection limit as negative values (i.e -1 is equivalent to < 1 mg/L or μ/L), considering that *PHREEQC* will disregard it. For variables as *pe* or *Eh*, this s clearly not applicable because redox state is a real number, that can be both with positive and negative sign. Non-measured values, or blank, are substituted by "nd" string;
- renaming of the original variable's names so that they comply with *PHREEQC* standards. For example, when sulfur is reported as sulfate in the original column, it is labeled as $S(6)$. In the input file, it will be flagged with "*as SO4*" in the SOLUTION-END data block;
- merging of laboratory pH columns with field pH measurements where no *in-situ* measurements were present (this happens in particular for older data);
- converting field's redox state measurement reported in mV to *pe* as required by *PHREEQC* corrected by *in-situ* temperature.

[2] (https://gis.epa.ie/GetData/Download, filename: EPA Groundwater Monitoring Data to End 2020 Circulation 26.05.21.xlsx)

Bash and SQL scripts are used to complete the data pre-processing and computation sequence presented in this section. Results and timing are based on speciation computations carried out on a MacBook with M1 (10 cores) ARM chip architecture and 16 Gb RAM.

6.2. Performances

The script `case_study.sh` consolidates all the *PHREESQL* capabilities into one, also assisting the reader in generating *PHREEQC* input format starting from a *.csv* file. As a matter of example and no loss of generality, we demonstrate the full capabilities of *PHREESQL* through *PHREESQLexe*.

Starting from the revised version of the EPA Ireland groundwaters dataset, we selected three subsets of data (`SHORT`, `MEDIUM` and `FULL`) as in Table 1, which were each processed using five *PHREEQC* thermodynamic databases (*llnl.dat, wateq4f.dat, phreeqc.dat, minteq.dat and sit.dat*). For all of them, data from a single ASCII file source of geochemical information is used to generate *.pqi*, *.met* and finally to create *.pqo* to be stored in `IN`, `META` and `OUT` folders respectively, as representative of input, metadata, and output folder. Examples are performed with these database because iterations to solve speciation-solubility equilibria lead to an expected variability on numerical results, and on execution performances. The main reasons rely on the diverse expression of the computation of the activity coefficients, on the number of master species, and on the number of aqueous species, solids and gases present on each database, see Lu et al. (2022).

As a matter of example, the *.pqo* files were generated by a single call to *PHREESQLexe*, running the following command line:

```
phreesqlexe --run_phreeqc -P llnl -i IN --o OUT
```

where input is stored in `IN` folder, the output will be stored in `OUT` folder, and the speciation computation will be based on a specific thermodynamic database, namely `llnl.dat`. The files thus generated in the `OUT` folders are labeled with the identical basename as in `IN` folder but different extensions, to facilitate shell programming. Parallelization of *PHREEQC* computation in *PHREESQL* enhances considerably the computation efficiency. By calling the speciation calculation directly from *PHREEQC*, the computing time is on average more than halved compared to running both *PHREESQL* as a serial process and a bash script, without requiring programming skills.

Once the results of the geochemical simulations are available, *PHREESQL* databases are generated by running the following command line:

```
phreesqlexe --fill_db -d PHREESQL_DB -i IN --o OUT --m META
```

where `PHREESQL_DB` is the database to be created and `IN`, `OUT` and `META` folders store inputs and outputs of *PHREEQC* and metadata respectively.

Table 1 shows that the performance of *PHREESQL* is highly dependent on sample size and thermodynamic database as expected. Analyzing Table 1 by column reveals the impact of the number of analyses presented in the dataset on computation time for each thermodynamic database. According to the results, using the `phreeqc.dat` database for computations results in an increase in performance from 1 (with the `SHORT` dataset) to 4.6 (with the `FULL` dataset), as demonstrated by the parallel/serial ratio. The same check can be performed by reading the table row by row. The final time is affected by the completeness of the thermodynamic database and the activity coefficient equation used, as demonstrated by the time resources required. The dataset with a higher number of analyses (`FULL`) exhibits a larger average increase in performance due to parallel computing. The ratio of serial to parallel performance has a significant increase from 4.6 to 4.9.

The number of successful analyses for each database as in Table 1 is slightly different from that in the data file *.csv*, due to some convergence errors attributed to incomplete chemical analyses and a resultant impossibility to solve the iterative computation.

The time needed to create a SQL database is largely dependent on the number of analyses and secondarily on the number of master species and equilibria are in the thermodynamics database. In this case, the differences due by thermodynamic database is much smaller and due to the length of the output ASCII file. When the computations of all thermodynamic databases are merged into the SQL database (`ALL`), the total time required is close to the sum of the time required by each individual database, indicating a linear scalability. After creating the relational database, the user is capable of browsing the results through either a command line or an external SQL graphical browser (not included in *PHREESQL*).

From Table 1, it can be shown what are main benefits:

- *PHREESQL* databases show a reduction in file size dimension compared to the same information in ASCII file format (with a ratio between 0.5 and 0.7), stored in one single file;
- the information is integrated into a standard file format and highly optimized for query;
- the programming skill is greatly reduced to a SQL query by avoiding text search on a large number of files (i.e. for `FULL` database the search should be extended to 3 times 14689 ASCII text files trying to maintain relationships among information in file of variable length).

Thanks to these features, complex queries may also be built in a more efficient way into a structured script and run by command line or through an external package or web server. To demonstrate the *PHREESQL* capability in querying *PHREEQC* speciation calculations stored in a *PHREESQL* database, we report some examples of data mining and use. In the following, we do not claim to explain the full geochemistry of Irish groundwater, but instead how a large dataset can be efficiently exploited by *PHREESQL*.

6.3. Data extraction and plotting

The following sections provide five examples illustrating how to query the *PHREESQL* database for data extraction and possible plotting by integrating *PHREESQL* with external graphical tools. The distribution includes scripts to replicate the entire set of examples, and the computational time for the database query is in real-time. The examples refer to *PHREESQL* databases generated by our case study, as described in Section 6.2, with specific reference to the `FULL` subset and the `wateq4f.dat` thermodynamic database when not otherwise specified.

6.3.1. Example 1: Groundwater classification

Surface and groundwater analyses can be plotted on well known diagrams such as those of Piper (Piper, 1945), Durov (Durov, 1948), Ficklin (Ficklin et al., 1992), Stiff (Stiff, 1951) or Langelier-Ludwig (Langelier and Ludwig, 1942) emphasizing the chemical features through a classification in geochemical types. Thermal waters can also be plotted in diagrams that rely on geothermobarometry suggesting equilibria with reservoir as in Giggenbach diagram (Giggenbach, 1988). Preprocessing data is necessary to ensure full consistency with space geometry (i.e. Euclidean, binary plots, simplex, triangular, or tetrahedral plots) and measurement units when building a classification plot. By categorizing groundwater based on various criteria and applying filters, users can effectively analyze and interpret data in a more flexible way. Such criteria can include rock type, county name, aquifer name, or time interval, and these categories can be plotted separately for better visualization and analysis. Additional filters can involve selecting specific time ranges or application of conditions related to electrical

Table 1
Performance of *PHREESQL*. [1]number of successful *PHREEQC* run stored in SQL database; [2] time in seconds for *PHREEQC* computations (parallel/serial); [3] time in seconds to fill SQL database; [4] SQL database size; [5] full suite of PHREEQC ASCII files size (input, output, and metadata); [6]ratio between SQL database and ASCII file size. [7]number of analyses that populate the *.csv* file. In the last column, computation data needed for the creation of a database comprehensive of all five *PHREEQC* aqueous databases are reported. Reference machine: Apple MacBook with 10 cores (M1 Apple chip) and 16 Gb RAM.

	phreeqc	*llnl*	*minteq*	*wateq4*	*sit*	ALL
SHORT (43)[7]	41[1]	41	41	41	41	205
	<1/<1[2] s	1/3 s	1/2 s	1/1 s	1/5 s	–
	0.2[3] s	0.4 s	0.3 s	0.3 s	0.3 s	1.3 s
	0.6[4] Mb	1.4 Mb	0.8 Mb	0.8 Mb	1.3 Mb	4.9 Mb
	1.2[5] Mb	2 Mb	1.3 Mb	1.4 Mb	1.8 Mb	7.7 Mb
	0.50[6]	0.70	0.61	0.57	0.72	0.63
MEDIUM (4355)	4319	4319	4319	4319	4319	21595
	9/42 s	63/332 s	31/139 s	18/79 s	96/510 s	–
	17 s	27 s	19 s	31 s	26 s	108 s
	64.4 Mb	189.4 Mb	107 Mb	112.5 Mb	165.2 Mb	646.1 Mb
	131 Mb	248 Mb	166 Mb	175 Mb	230 Mb	950 Mb
	0.49	0.75	0.64	0.64	0.72	0.68
FULL (14690)	13598	13598	13598	13598	13598	67990
	28/129 s	200/1073 s	97/457 s	54/248 s	295/1468 s	–
	44 s	75 s	56 s	57 s	68 s	297 s
	204.6 Mb	502.4 Mb	299.4 Mb	305.1 Mb	452.5 Mb	1.8 Gb
	407 Mb	704 Mb	494 Mb	506 Mb	708 Mb	2.75 Gb
	0.50	0.71	0.61	0.60	0.63	0.65

balance or concentration ranges. This helps to refine the dataset and focus on specific aspects of groundwater data.

The SQL script `LL.sql`, called by the `LL.sh` bash script, filters groundwater samples from the input *PHREESQL* database. The filter is based on the county's name, allowing users to retrieve data from specific counties without applying additional compositional or temporal filters. The output consists of three ASCII files, one for the whole dataset and two for a pair of Irish counties (Kilkenny and Donegal, arbitrarily chosen). Fig. 2 shows an example of LL plot (Langelier and Ludwig, 1942) given the filter output. The diagram is similar to the Piper diagram plotting only major cations and anions. By grouping Cl, SO_4, and HCO_3 and K, Ca and Mg and taking into account the following relations:

$$\%Ca + \%Mg = 50 - (\%Na + \%K) \quad (1)$$

$$\%Cl = 50 - (\%HCO3 + \%SO4) \quad (2)$$

cold water chemical types can be assigned to each analysis. One of the advantages of using an LL plot is that mixing lines are straight, and vertices can indicate the compositions of natural salts such as calcite, anhydrite, and halite. Trends moving closer or further away from these vertices can suggest dissolution or precipitation of these salts. Simultaneously, there is a partial loss of information, avoiding the discrimination between grouped ions in relative equivalent concentration.

The plot highlights that the waters in Kilkenny County are divided into two distinct clusters, with the first group consisting of $CaMg - HCO_3$ waters and the second group having relevant $NaCl - sulfate$ concentrations. By not presuming to draw conclusions, interested users can be assisted in understanding variability and grouping, due to:

- seasonal oscillations;
- spatial variability because water-rock interaction due by hosting rocks;
- incompleteness about alkalinity or total carbon evaluation.

By customizing the query, and looping over all the 26 counties, the LL diagram shows atomic information about each county.

6.3.2. Example 2: Water-rock interaction

Groundwater equilibria can be better understood through speciation-solubility calculations, which can also represent the relationships with host rocks and the tendency of the equilibrium state to change. The

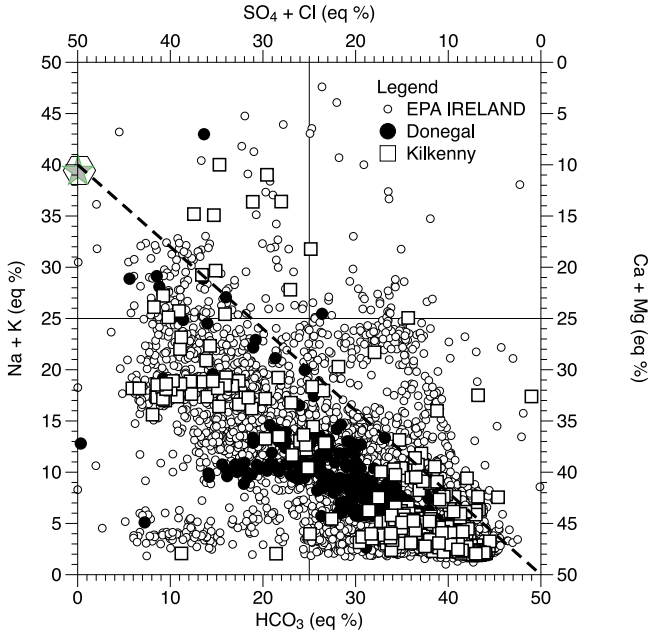

Fig. 2. Example 1. Langelier-Ludwig classification diagram with data extracted with the script *LL.sql* and plotted with Datagraph©. The line represent the dilution line between seawater (hexagon) and $Ca - HCO_3$ water. Green star identify a coastal rain composition, hexagon a standard seawater.

solution saturation index is used to assess that condition and it is defined as:

$$SI = log \frac{K_{IAP}}{K_{sp}} \quad (3)$$

where K_{sp} and K_{IAP} are the equilibrium constant and the ionic activity product, respectively. Using the T and P of interest expressed in the original dataset, *PHREEQC* computes both K of the reaction used to describe the phenomena and reported into thermodynamic databases.

In Fig. 3, waters of Kilkenny and Donegal counties are reported by extracting data limiting the *pH* range (6.35–10.33) where HCO_3^- anion is the predominant species at T, P standard condition (25° and 1 atm), superimposed to the full Irish dataset. The query uses a constraint for *pH*, and water equilibria are calculated at temperature and pressure from the EPA dataset. This leads to a plot showing the solution's

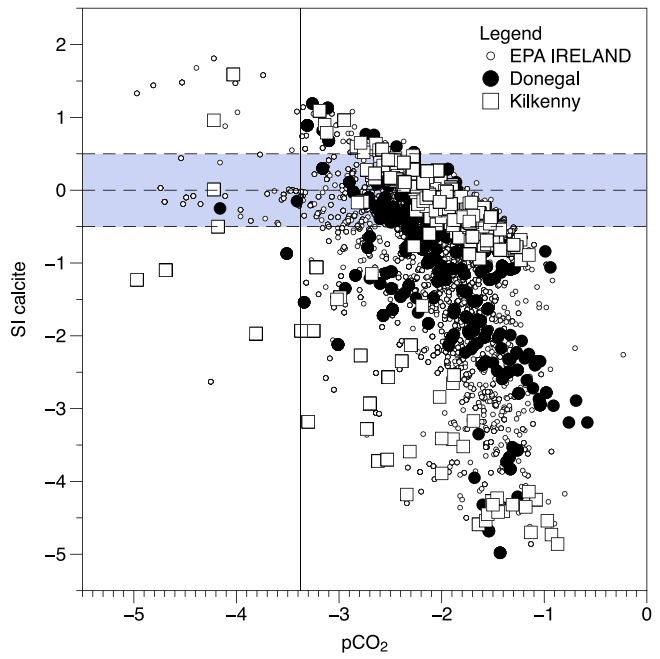

Fig. 3. Example 2. Calcite saturation index vs pCO_2 with data extracted with the script *sat_indx.sql* and plotted with Datagraph©. Water can be considered in equilibrium with calcite in grayed zone, undersaturated below −0.5, oversaturated above 0.5. The pCO_2 value of −3.38 (416 ppmv) is the actual carbon dioxide concentration in atmosphere.

saturation index relative to calcite and total dissolved inorganic carbon or TDIC (as $-log$ of the CO_2 fugacity). The relationship between the dissolution in water of carbon dioxide and the dissolution of calcite, as described in the carbonate system reactions (Eq. (6)), is highlighted by the trends shown in the figure.

$$CO_{2(aq)} + 2H_2O \leftrightarrow H_2CO_{3(aq)} \tag{4}$$

$$H_2CO_{3(aq)} \leftrightarrow HCO^-_{3(aq)} + H^+_{(aq)} \tag{5}$$

$$CaCO_3 + H^+_{aq} \leftrightarrow Ca^{+2}_{aq} + HCO^-_{3(aq)} \tag{6}$$

A clustering is still visible as in Fig. 2 implying that over the entire time periods reported in the EPA file, some wells in Kilkenny county have shown different saturation conditions. There is usually undersaturation, although most available waters remain close to equilibrium with calcite. The water that was previously recognized as clustered is presumably the same in both group. According to this suggestion, calcite undersaturation can be a consequence of a water composition change, with a decrease in calcium and magnesium in favor of sodium and chloride and an increase of dissolved CO_2.

Although waters in Donegal county are less dispersed during this period, they are still under-saturated, which may be due to the interaction with the granitic substrate and sandstones, mudstone and greywackes. These rocks regulate equilibria, which are mostly dominated by alumina-silicate phase minerals rather than carbonate.

6.3.3. Example 3: Eh-pH diagram by external software

In this section, we illustrate the process of calling an external package, specifically *PhreePlot*, to create a graphical output. This involves first extracting the original data from the SQL database and convert the expression of the redox state from one scale (pe) to another, as that of E_h. Measurement of labile physical–chemical parameters (as pH) and redox state (as Eh) is mandatory for a reliable groundwater speciation calculation aiming to describe the in-situ water natural condition. The predominance plot is known also as Pourbaix plot (Pourbaix, 1966) or more generally $E_h - pH$ diagram. It can be used to discriminate water environment respect to water stability field (Baas Becking et al., 1960) based on field measurement. By specifying a chemical system, including its pressure and temperature parameters, and performing a speciation calculation while iterating over a range of redox and pH values, we can determine the fields where aqueous species predominate.

The $E_h - pH$ diagram shown in Fig. 4 represents the predominance fields of aqueous species of a As-O-C-S-H system at standard pressure and temperature (25 °C and 1 bar) and As=40 μg/l. The arsenic has been chosen as an example because it represents a risk in drinking waters in Ireland (McGrory et al., 2017, 2018, 2021a). Additionally, it is sensitive to be released by change in environmental parameters as alkalinity, dissolved organic matter, sulfate or chemical-physical parameters as in Smedley and Kinniburgh (2002).

The plot represents a snapshot at a fixed pressure and temperature, and its validity is virtually limited to those conditions. At different thermobaric ambient values, the isoactivity boundaries are expected to shift parallel to themselves expanding or reducing predominance fields. Temperatures of water samples stored in the EPA SQL database are generally lower than standard, reflecting groundwater environment, and the pe was referred to the *in-situ* temperature. Although it is an approximation, plotting a set of labile parameter pairs (as $E_h - pH$) on a fixed T-P system can still provide information about arsenic speciation. The script `03_eh_ph.sh` firstly extracts $pe - pH$ couples from the input *PHREESQL* database for both Kilkenny and Donegal counties, as previously done, with a single constraint on charge balance within a range of ±10%. The redox parameter $pe = -log[e^-]$, which has thermodynamic meaning is converted to E_h in [V] to be compared with field measurement with the equation:

$$E_h = pe\ 2.303 \left[\frac{RT}{F}\right] \tag{7}$$

where R is the gas constant, T is temperature in Kelvin as reported in SQL database, and F is the Faraday constant. Passing the output of the constrained SQL query to the *PhreePlot* template input file enables the creation of a diagram that represents the two counties separately. This predominance diagram, taking into account previous approximations, provides a comprehensive description of the waters with respect to the $As(III)/As(V)$ redox pair, regardless of the total As concentration. Water samples from Kilkenny County are found in a field that is almost exclusively characterized by arsenite as $As(III)$ species. The pattern of water samples from Kilkenny is distributed ranging mainly over pH with a restricted redox variation, in the field of $As(V)$ species. Data processing of constrained dataset can be beneficial in gaining insights into health protection and remediation activities, as arsenites and arsenates have different toxicity and can be treated differently.

6.3.4. Example 4: Speciation results database dependency

The quality of water analysis and the aqueous database used have a significant impact on the outcome of speciation calculation. The first feature tends to ensure that the geochemical system is sufficiently described in terms of elements and guarantees a good approximation of electroneutrality. The second feature describes the consistency of reactions between aqueous species, dissolved gases or solids in heterogeneous systems, ending in species activity calculation. A full internal consistency of an aqueous database is a “nirvana” of geochemists (van der Lee and Lomenech, 2004) although *SUPCRT* (Johnson et al., 1992) is one of older databases with this feature, a large number of thermodynamic databases are available, with a more or less restricted field of application (Lu et al., 2022). *PHREEQC* is released with an increasing number of databases, to which are added databases released independently, by optimizing chemical system, P,T range, completeness and activity coefficient equation. Due to this richness of databases, and the iterative method of computation of equilibria users are facing non-unique results (Lu et al., 2022). This section shows how *PHREESQL* can be used to compare results by collecting computation done with the 5 thermodynamic database on the same dataset. The script `04_SI_compare_ALL.sh` extracts (by running 5 serial queries) the

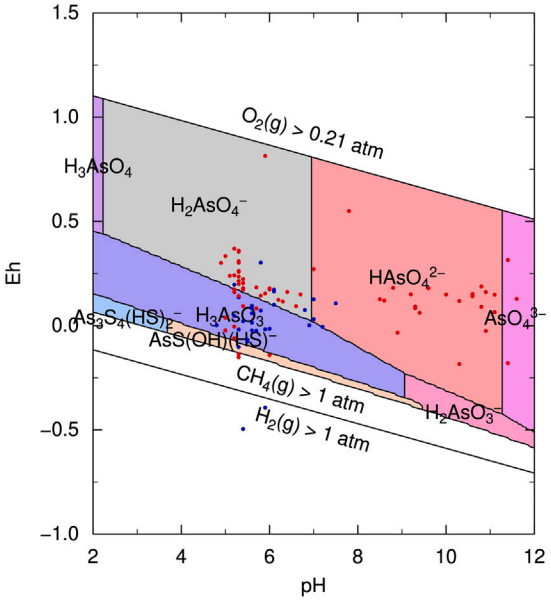

Fig. 4. Example 3. Eh-pH diagram of As-O-C-S-H system at 25 °Celsius using *PhreePlot* as external package by `04_ehph.sh` shell script. Blue symbols, Kilkenny county, red symbols Donegal county.

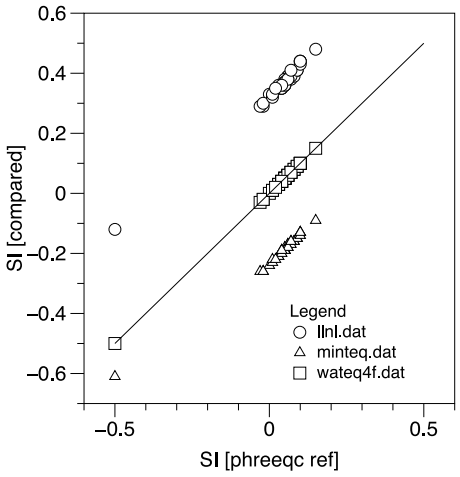

Fig. 5. Example 4. Solution saturation indices comparison respect to chalcedony for a single sampling site among coexisting computations in SQL database `FULL_ALL.db`. `phreeqc.dat` database is used as reference datum.

saturation index of chalcedony for a single sampled site (EPA code: IE_EA_G_0010_2100_0009) over decades. Despite the limited concentration of silicon in surface and groundwater, it has a significant impact on environmental evolution. It is a product of weathering of alumino-silicate minerals, and it is recycled as components of secondary minerals or as the main material for diatoms (Neal et al., 2005). Variations in pH, temperature, or CO_2 fugacity in the atmosphere or flowing water can be the cause of multi-temporal concentration variability in waters, thus inducing change in equilibria in dissolution-precipitation processes and kinetics. In Smedley and Kinniburgh (2023) can be shown that SiO_2 polymorphs can adsorbs or traps U, which is why it is crucial to assess the saturation state of a solution in relation to a specific phase. In Fig. 5, the `phreeqc.dat` default database is chosen as a reference and the other are compared to it. It can be observed that `wateq4f.dat` is perfectly consistent with `phreeqc.dat` as the latter is derived and less complete, suggesting a quasi-perfect saturation with SI_{ch} close to 0. Two databases as `llnl.dat` and `minteq.dat` provide a contrasting result with SI values of opposite sign. Although SI values are in the range ± 0.5, which is often considered to be a nearly saturated condition, it is suggested that depending on the considered model, it is possible to develop under- or over-saturation scenarios. Due to the fact that in the `sit.dat` database, the chalcedony is not present it is not possible to report the saturation data set in Fig. 5.

6.3.5. Example 5: Geochemical mapping

The widespread use of GIS tools and of online spatial databases requires that information about sample's locations become essential and a strict linkage with environmental data. By generating spatial–temporal datasets derived from speciation calculations, transport-reaction modeling, or spatial distribution analysis, we can obtain an approximate depiction of reality, fostering a deep comprehension of the environment. Through *PHREESQL*, information about locations reported in original *.csv* files is managed through *.met* files. In this section, we show how easily geochemical outcome of *PHREEQC* can be extracted together positional attribute, converting *on-the-fly* a CSR to another to be compliant with a geographical database. The example is intended to demonstrate how to prepare geochemical data (georeferenced) to be used by external software for mapping through a stochastic method, although details are not provided since they are out of the scope of this work.

Although *PHREESQL* is not able to finalize variography, its integration with external geostatistical packages (i.e. GSLIB (Deutsch and Journel, 1998), GSTAT (Gräler et al., 2016), or MUSE (Miola et al., 2022)) enables the computation of experimental variograms and of their model (see Fig. 6).

Spatial covariance defined by experimental variograms is necessary by kriging algorithms or stochastic simulation with the aim to calculate the spatial distribution of a variable and its uncertainty. Mapping can then be done both for a raw elemental concentration, or for a derived one as aqueous species or even for mineral saturation indices of solutions stored in a SQL database generated by *PHREESQL*. Bearing in mind that countries may have different co-existing coordinate systems for mapping and for distributed digital file format, *PHREESQL* can convert/transform coordinates from an EPSG code to another. Geological Survey of Ireland distributes its vector or raster geospatial files under EPSG:29902 (TM65/Irish Grid Ireland) or EPSG:2157 (IRENET95/Irish Transverse Mercatore.) EPA groundwater analyses coordinates uses TM65, but associated maps (as bedrock or aquifer composition) use IRENET95. By `05_cs2cs.sh`, the user can check how to export into a spatially converted database, a set fo solution saturation indices respect to calcite for a specific period (i.e. July 1st to Aug 31st 2012). The selected time period allows to have the largest number of analyses over the entire national territory, without duplicates neglecting subsoil water as described in EPA file. A variogram is computed experimentally as in Eq. (8), modeled by permissible models, before a stochastic simulations is performed by external packages (in our example GSLIB was used).

$$\gamma(h)_{SI^{cc}} = \frac{1}{2N}\sum (SI_i^{cc} - SI_j^{cc})^2 \tag{8}$$

$$g(h)_{SI^{cc}} = 1.5\frac{h}{a} - 0.5\left(\frac{h}{a}\right)^3 \tag{9}$$

In Eq. (8), $\gamma(h)_{SI^{cc}}$ is the experimental semivariogram, N is the number of sample pairs, joined by vectors, over all analyses falling in user-defined tolerances around h, and $SI_i^{cc} - SI_j^{cc}$ is the difference between the saturation indices of the vector tail and head, indicated as i and j respectively, processed as regionalized random variables. Eq. (9) represents a (positive defined) spherical model used in the kriging matrix solution. In Fig. 6, the experimental variogram is depicted by white circles modeled up to an asymptotically value of the regional variance of 0.7 by a spherical model, with a nugget effect of 0.26 (about 37% of total variance). Once a covariance model about SI_{cc} is

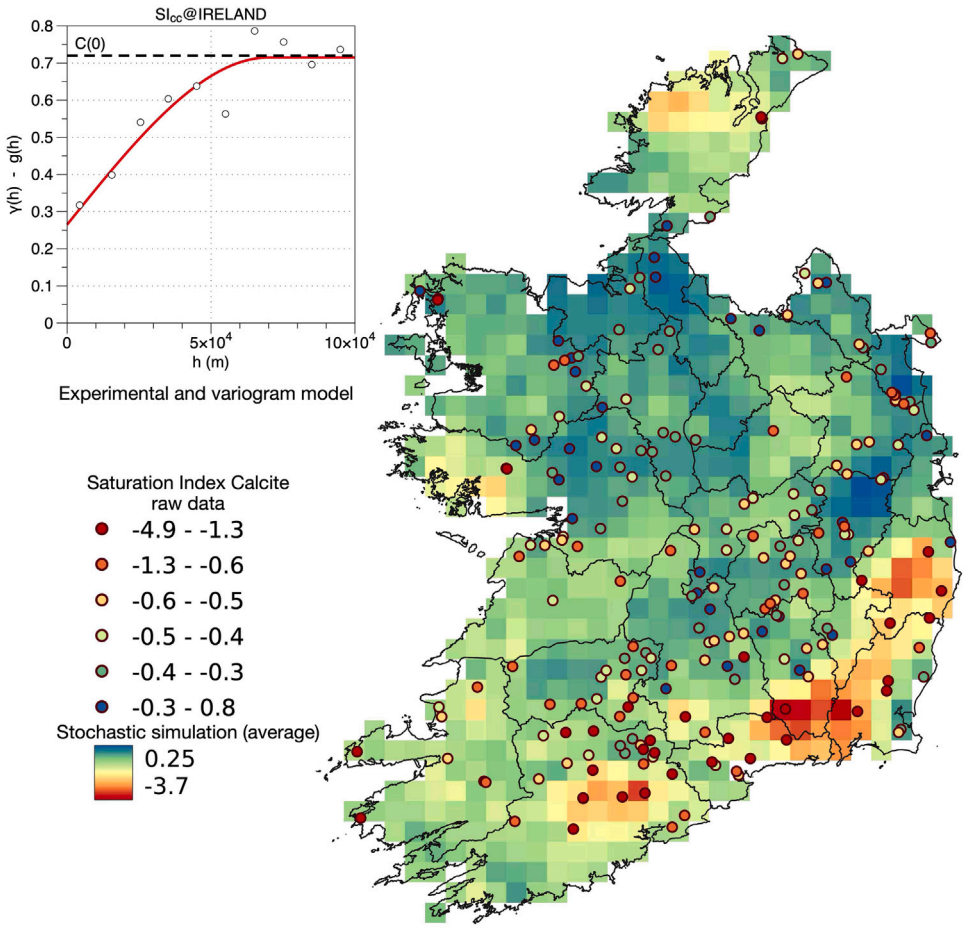

Fig. 6. Example 5. Stochastic image of average value of solution's saturation index respect to calcite of groundwater by using conditional Sequential Gaussian Simulation GSLIB and visualized with GRASS. Each cell has a spatial resolution of 10 km. The map show counties from epa.gov.ie repository.

obtained (see Eq. (9)), a conditional stochastic simulation of solution's calcite saturation index with a resolution of 10 km can be performed, as visible in Fig. 6. Stochastic simulations are the main tool to describe uncertainty of a random variable for each computational domain by reconstructing the probability density function of a random variable. In Fig. 6b, the average SI_{cc} is plotted. Inspecting Fig. 6, the distribution of the saturation index for calcite (although over a coarse grid) suggests a well defined undersaturated zone, related to paleozoic deep marine and graywackes formations.

In this particular case once spatial statistics has been computed, results were imported into a GIS as GRASS (Neteler et al., 2012; GRASS Development Team, 2022) to be further processed and to produce Fig. 6b.

Running a loop over other variables, as dissolved species or gas fugacities, ranging on different time lags, users can exploit the surface or groundwater geochemistry of a large and dense database over the temporal dimension.

7. Conclusions

PHREESQL is designed to efficiently store large geochemical water datasets for speciation-solubility calculations. It enables to permanently store and browse *PHREEQC* output in a SQL database. The code is released as an open source library and as a command line application. Specifically, the toolkit includes both a code library (PHREESQLib) and a standalone application (PHREESQLexe). *PHREESQLib* is a C++ header-only library: it is easily usable without compilation, as third-party software developers can include it into their projects. *PHREESQLexe* is a standalone C++ application leveraging all the features of *PHREESQLib*, primarily generating relational SQL databases storing all *PHREEQC* calculations.

Once the mandatory data file pair (as *.pqi* and *.met*) is created, PHREESQL can efficiently perform aqueous speciation calculations, benefiting from parallel computation techniques to reduce the simulation computation time compared to the traditional usage of *PHREEQC*. This enables *real-time* processing of speciation calculation in experiments where parameters are acquired by sensors, both in fiend and laboratory settings.

The SQL database is built following the *PHREEQC* output sections adding a metadata table with descriptive and georeferenced data. Embedding the PROJ library, coordinates can be converted and/or transformed *on the fly* while creating a SQL database, duplicating a new database with coordinates with a different EPSG code, or exporting coordinates in an ASCII file. This option empowers the capability to homogenize the geospatial features about analysis metadata framed in heterogeneous CRSs into a single geographical system. This functionality allows the integration with external GIS packages and full spatial processing of all *PHREEQC* variables.

PHREESQL has a robust set of rules for internal coherence checks, preventing data duplication. It allows the coexistence of the same speciation-solubility computation if analyses are computed with diverse *PHREEQC* thermodynamic databases. The standard structure of a SQL database created by *PHREESQL* can be updated by external commands, enabling modification or creation of new tables, preferably associated with the `METADATA` table, using SQL commands.

PHREESQL is ready to be embedded into third-party geochemical applications. The capability to embed a such large quantity of information related to surface or groundwater speciation-calculation into a

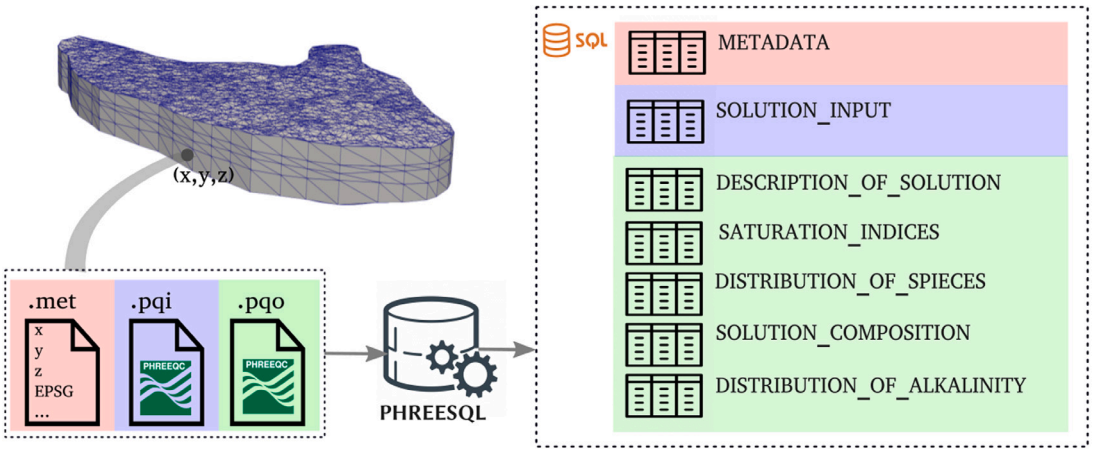

Fig. 7. The joint exploitation of *PHREESQL* and 3D unstructured meshes is useful for reactive transport model. Each point in the unstructured mesh can be described as a triplet of files encoding metadata (*.met*), *PHREEQC* input (*.pqi*) and *PHREEQC* output (*.pqo*) respectively. Such a description enables the possibility to exploit *PHREESQL* to efficiently store and query geochemical simulation results. Colors are used to show the mapping between *PHREESQL* input and *PHREESQL* database structure.

compact and efficient SQL data format makes *PHREESQL* an attractive tool for the implementation also in reaction-transport modeling, GIS and web applications.

The library has been written with the aim to be easily integrated in third-party codes and to support the building of extended databases related to 3D unstructured meshes as a support of reaction-transport (RT) models (see Fig. 7). The structure of a *PHREESQL* database can be easily extended by adding dedicated tables to encode the evolution of reaction-transport models in both space and time.

7.1. Limitations and future works

At its present stage, *PHREESQL* is limited to performing speciation calculations generated exclusively by *PHREEQC*. It supports only certain types of inputs and outputs, requiring a single "SOLUTION-END" block and optionally the "print alkalinity" clause in the input file. However, these limitations are expected to be addressed in forthcoming releases, with the potential to accommodate a broader spectrum of *PHREEQC* options as "EQUILIBRIUM_PHASES", or "REACTION_TEMPERATURE" inputs and outputs from other speciation calculation tools, such as EQ3/6 (Wolery and USDOE, 2010) and PFLOTRAN (Hammond et al., 2019).

The current distribution includes a comprehensive case study along with accompanying scripts designed to illustrate the proper creation of input datasets for *PHREESQL* using real data. These scripts are tailored specifically to our case study, and there is ongoing effort to adapt them for more generalized use with other datasets.

From the usability point of view, *PHREESQL* lacks of a graphical interface. Consequently, its usage may be less intuitive and user-friendly, particularly for individuals who are not comfortable with command-line interactions or have limited programming skills. Some tasks may involve the execution of lengthy and intricate commands with a multitude of options, flags, and parameters. This level of complexity increases the likelihood of errors, particularly when dealing with challenging tasks considering nested constraints. Additionally, proficiency in SQL is necessary to query the generated databases and extract specific subsets of data. To enhance *PHREESQL*'s usability and its utility, future endeavors should focus on mitigating these issues, ultimately striving to provide visual feedback and a more interactive experience for users.

Code availability section

Name of the code/library: ***PHREESQL***

Contact: e-mail and phone number: **Daniela Cabiddu**, E: **daniela.cabiddu@cnr.it**, P: **+39 010 6475696**

Hardware requirements: **None**

Program language: **C++**

Software required: **SQlite**, **CMake and C++ compiler** (for building purposes)

Program size: source code library 228 K

The source codes are available for downloading at the link: https://github.com/DanielaCabiddu/PHREESQL

CRediT authorship contribution statement

Marino Vetuschi Zuccolini: Funding acquisition, Methodology, Conceptualization, Validation, Writing – original draft, Writing – review & editing. **Daniela Cabiddu:** Methodology, Conceptualization, Software, Writing – original draft, Writing – review & editing. **Simone Pittaluga:** Conceptualization, Software, Writing – original draft, Writing – review & editing.

Declaration of competing interest

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

Data availability

Data are distributed freely by EPA Ireland under CCA 4.0.

Acknowledgments

This research was supported by funding from the European Union - NextGenerationEU and the Ministry of University and Research (MUR) under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.5, for the project "RAISE - Robotics and AI for Socio-economic Empowerment" (ECS00000035). The authors would also like to thank David Kinniburgh and all the anonymous reviewers for their valuable comments and suggestions on the paper.

Table 2
Command line options and arguments to run *PHREESQLexe*.

Action	Option [<argument>]	Option (short)	Description
--run_phreeqc			Run speciation calculation.
	--phreeqc_db <db_path>	-P	Path of *PHREEQC* database to be used for the calculation.
	--in_folder <folder_name>	-i	Path of the folder containing *PHREEQC* inputs (*.pqi*).
	--out_folder <folder_name>	-o	Path of the folder where *PHREEQC* output (*.pqo*) will be saved.
--fill_db			Create a new database or update an existing one.
	--database <db_path>	-d	Reference to the database to be created/updated.
	--in_folder <folder_path>	-i	Path of the folder containing *PHREEQC* inputs (*.pqi*).
	--out_folder <folder_path>	-o	Path of the folder containing *PHREEQC* outputs (*.pqo*).
	--meta_folder <folder_name>	-m	Path of the folder containing metadata files (*.met*).
	--overwrite		If an input analysis is already in the database, overwrite it.
--export_input			Export *PHREEQC* input (*.pqi*) from the database.
--export_output			Export *PHREEQC* output (*.pqo*) from the database.
--export_metadata			Export *PHREEQC* metadata (*.met*) from the database.
--export_all			Export *PHREEQC* input, output and metadata from the database.
	--database <db_name>	-d	Reference to the input *PHREESQL* database.
	--export_folder	-F	Path of the output folder.
	--export_id	-I	The ID of the single analyses to be exported.
	--export_list_ids	-L	Path of a text file listing the IDS of the analyses to be exported.
--epsg_convert			Export the database by converting the coordinates into a desired EPSG.
	--database	-d	Reference to the input *PHREESQL* database.
	--epsg	-e	Output EPSG.
	--out_filename	-f	Path of the output file (.csv).
	--out_table	-t	Name of the output table to be created in the input *PHREESQL* database.
	--out_database	-b	Name of the output *PHREESQL* database having converted EPSG.

Appendix A. Metadata file structure

The metadata file structure is designed for the purpose of processing *ex-situ* or *in-situ* real-time data. It has a very basic structure reporting information on spatial and temporal location:

- JOB TYPE: Name of Project;
- SURVEY: Name of survey, or relevant geographical information (e.g. Region for Italy, County for Ireland or Department for France);
- SITE: neighboring place-name of the sampling site;
- DATE: survey date in yyyy/mm/dd format;
- DATABASE: *PHREEQC* thermodynamic database used in computation;
- PHREEQC_VERSION: version of the *PHREEQC* used in computation;
- RUN_NUMBER: to be used to identify the sequential steps of iterative processes (e.g. simulations);
- SAMPLE_NAME: original dataset or report sample ID;
- INPUT_FILE: full name with *PHREEQC* extension of input file, present in scratch directory;
- COORD_X, COORD_Y, COORD_Z: coordinates of sampled point consistent with the EPSG field;
- EPSG: European Petroleum Survey Group code for geodetic parameters;
- TIMESTAMP: UNIX timestamp of sampled point, mandatory for surveys with a high rate of sampling.

Appendix B. Command line options

PHREESQLexe allows exploiting the main functionalities of *PHREESQLib* by command line. Table 2 lists and describes both arguments and options of the tools, to clarify how to use it.

References

Appelo, C., Rolle, M., 2010. PHT3D: A reactive multicomponent transport model for saturated porous media. Groundwater 48, 627–632. http://dx.doi.org/10.1111/j.1745-6584.2010.00732.x.

Baas Becking, L., Kaplan, I., Moore, D., 1960. Limits of the natural environment in terms of pH and oxidation-reduction potentials. J. Geol. 68, http://dx.doi.org/10.1086/626659.

Berretta, S., Cabiddu, D., Pittaluga, S., Mortara, M., Spagnuolo, M., Vetuschi Zuccolini, M., 2018. Adaptive environmental sampling: The interplay between geostatistics and geometry. In: Livesu, M., Pintore, G., Signoroni, A. (Eds.), Smart Tools and Applications in Graphics. pp. 133–140.

Blanc, P., Lassin, A., Piantone, P., Azaroual, M., Jacquemet, N., Fabbric, A., Gaucher, E., 2012. Thermoddem: A geochemical database focused on low temperature water/rock interactions and waste materials. Appl. Geochem. 27, 2107–2116. http://dx.doi.org/10.1016/j.apgeochem.2012.06.002.

Caccia, M., Ferretti, R., Odetti, A., Bruzzone, G., Spagnuolo, M., Mortara, M., Berretta, S., Cabiddu, D., Pittaluga, S., Vetuschi Zuccolini, M., Brignone, L., 2019. http://dx.doi.org/10.1109/OCEANSE.2019.8867568.

Chandra, R., Dagum, L., Kohr, D., Menon, R., D. Maydan, J.M., 2001. Parallel Programming in OpenMP. Morgan kaufmann.

Charlton, S.R., Macklin, C.L., Parkhurst, D.L., 1997. PHREEQCI—A graphical user interface for the geochemical computer program PHREEQC. Water-Resour. Investig. Rep. 9, 7–4222.

Charlton, S.R., Parkhurst, D.L., 2011. Modules based on the geochemical model PHREEQC for use in scripting and programming languages. Comput. Geosci. 37, 1653–1663. http://dx.doi.org/10.1016/j.cageo.2011.02.005.

Charlton, S., Parkhurst, D., Appelo, C., with contributions from D. Gillespie for Chipmunk BASIC, Cohen, S., Hindmarsh, A., Serban, R., Shumaker, D., A.G. Taylor for CVODE/SUNDIALS, 2022. Phreeqc: R interface to geochemical modeling software. URL https://CRAN.R-project.org/package=phreeqc. R package version 3.7.4.

Deutsch, C., Journel, A., 1998. GSLIB: Geostatistical Software Library and User's Guide, second ed. Oxford, p. 369.

Durov, S., 1948. Natural waters and graphic representation of their composition. In: Dokl Akad Nauk SSSR. 59, (3), pp. 87–90.

EPA-IE, 2023. Water and data. URL https://gis.epa.ie/GetData/Download. (Accessed 13 January 2023).

EPA-USA, 2023. Water and data. URL https://www.epa.gov/waterdata/water-quality-data-download. (Accessed 13 January 2023).

Federal Ministry Republic of Austria, 2023. Water and data. URL https://info.bml.gv.at/en/topics/water/water-and-data-wisa.html. (Accessed 13 January 2023).

Ficklin, W., Plumlee, G., Smith, K., McHugh, J., 1992. Geochemical classification of mine drainages and natural drainages in mineralized areas. In: International Symposium on Water-Rock Interaction. pp. 381–384.

Geological Survey of Ireland, 2023. Water and data. URL https://www.gsi.ie/en-ie/data-and-maps/Pages/Groundwater.aspx. (Accessed 13 January 2023).

Giffaut, E., Grivé, M., Blanc, P., Vieillard, P., Colàs, E., Gailhanou, H., Gaboreau, S., Marty, N., Madé, B., Durob, L., 2014. Andra thermodynamic database for performance assessment: ThermoChimie. Appl. Geochem. 49, 225–236. http://dx.doi.org/10.1016/j.apgeochem.2014.05.007.

Giggenbach, W., 1988. Geothermal solute equilibria. Derivation of Na-K-Mg-Ca geoindicators. Geochim. Cosmochimica Acta 52, 2749–2765.

Gräler, B., Pebesma, E., Heuvelink, G., 2016. Spatio-temporal interpolation using gstat. R J. 8, 204–218, URL https://journal.r-project.org/archive/2016/RJ-2016-014/index.html.

GRASS Development Team, 2022. Geographic resources analysis support system (GRASS GIS) software, version 8.2. Open Source Geospatial Foundation. URL https://grass.osgeo.org.

Guo, B., Hong, Y., Qiao, G., Ou, J., 2018. A COMSOL-PHREEQC interface for modeling the multi-species transport of saturated cement-based materials. Constr. Build. Mater. 187, 839–853. http://dx.doi.org/10.1016/j.conbuildmat.2018.07.242.

Hammond, G.E., Lichtner, P.C., Lu, C., Mills, R.T., 2019. PFLOTRAN: Reactive flow & transport code for use on laptops to leadership-class supercomputers. In: Groundwater Reactive Transport Models. http://dx.doi.org/10.2174/9781608053063112010l, URL https://www.osti.gov/biblio/1567668.

Hipp, R.D., 2020. SQLite. URL https://www.sqlite.org/index.html.

Johnson, J., Oelkers, E., Helgeson, H., 1992. Supcrt92: A software package for calculating the standard molal thermodynamic properties of minerals, gases, aqueous species, and reactions from 1 to 5000 bar and 0 to 1000°C. Comput. Geosci. 18 (7), 899–947.

Kinniburgh, D., Cooper, D., 2011. PhreePlot - creating graphical output with phreeqc. URL https://www.phreeplot.org/PhreePlot.pdf.

Korrani, A., Sepehrnoori, K., Delshad, M., 2015. Coupling IPhreeqc with UTCHEM to model reactive flow and transport. Comput. Geosci. 82, 152–169. http://dx.doi.org/10.1111/j.1745-6584.2010.00732.x.

Langelier, W., Ludwig, H., 1942. J. Am. Water Works Assoc. 34 (3), 335–352.

Leal, A., Kulik, D., Smith, W.R., Saar, M.O., 2017. An overview of computational methods for chemical equilibrium and kinetic calculations for geochemical and reactive transport modeling. Pure Appl. Chem. 89 (5), 597–643. http://dx.doi.org/10.1515/pac-2016-1107.

Lothenbach, B., Kulik, D.A., Matschei, T., Balonis, M., Baquerizo, L., Dilnesa, B., Miron, G.D., Myers, R.J., 2019. Cemdata18: A chemical thermodynamic database for hydrated portland cements and alkali-activated materials. Cem. Concr. Res. 115, 472–506. http://dx.doi.org/10.1016/j.cemconres.2018.04.018.

Lu, P., Zhang, G., Apps, J., Zhu, C., 2022. Comparison of thermodynamic data files for PHREEQC. Earth-Sci. Rev. 225, 103888. http://dx.doi.org/10.1016/j.earscirev.2021.103888.

Managed by the International Association of Oil and Gas Producers IOGP, 2023. EPSG, geodetic parameters dataset. URL https://epsg.org/home.html. (Accessed 5 April 2023).

McGrattan, K., Hostikka, S., Floyd, J., McDermott, R., Vanella, M., 2020. Fire dynamics simulator user's guide. https://github.com/firemodels/fds/releases/download/FDS6.7.5/FDS_User_Guide.pdf.

McGrory, E.R., Brown, C., Bargary, N., Williams, N.H., Mannix, A., Zhang, C., Henry, T., Daly, E., Nicholas, S., Petrunic, B.M., Lee, M., Morrison, L., 2017. Arsenic contamination of drinking water in Ireland: A spatial analysis of occurrence and potential risk. Sci. Total Environ. 579, 1863–1875. http://dx.doi.org/10.1016/j.scitotenv.2016.11.171, URL https://www.ncbi.nlm.nih.gov/pubmed/27932216. McGrory, Ellen R Brown, Colin Bargary, Norma Williams, Natalya Hunter Mannix, Anthony Zhang, Chaosheng Henry, Tiernan Daly, Eve Nicholas, Sarah Petrunic, Barbara M Lee, Monica Morrison, Liam eng Netherlands 2016/12/10 Sci Total Environ. 2017 Feb 1;579:1863-1875. doi: 10.1016/j.scitotenv.2016.11.171. Epub 2016 Dec 6..

McGrory, E., Henry, T., Conroy, P., Morrison, L., 2021a. Occurrence, geochemistry and speciation of elevated arsenic concentrations in a fractured bedrock aquifer system. Arch. Environ. Contam. Toxicol. 81 (3), 414–437. http://dx.doi.org/10.1007/s00244-021-00887-3, URL https://www.ncbi.nlm.nih.gov/pubmed/34519866. McGrory, Ellen Henry, Tiernan Conroy, Peter Morrison, Liam eng 2021/09/15 Arch Environ Contam Toxicol. 2021 Oct;81(3):414-437. doi: 10.1007/s00244-021-00887-3. Epub 2021 Sep 14..

McGrory, E., Henry, T., Morrison, L., 2021b. Occurrence, geochemistry and speciation of elevated arsenic concentrations in a fractured bedrock aquifer system. Arch. Environ. Contam. Toxicol. 81, 414–437. http://dx.doi.org/10.1007/s00244-021-00887-3.

McGrory, E., Holian, E., Alvarez-Iglesias, A., Bargary, N., McGillicuddy, E.J., Henry, T., Daly, E., Morrison, L., 2018. Arsenic in groundwater in South West Ireland: Occurrence, controls, and hydrochemistry. Front. Environ. Sci. 6, http://dx.doi.org/10.3389/fenvs.2018.00154, URL https://www.frontiersin.org/articles/10.3389/fenvs.2018.00154.

Miola, M., Cabiddu, D., Pittaluga, S., Vetuschi Zuccolini, M., 2022. MUSE: Modeling uncertainty as a support for environment. In: Cabiddu, D., Schneider, T., Allegra, D., Catalano, C.E., Cherchi, G., Scateni, R. (Eds.), Smart Tools and Applications in Graphics - Eurographics Italian Chapter Conference. The Eurographics Association, http://dx.doi.org/10.2312/stag.20221265.

Mosai, A., Tokwana, B.C., Tutu, H., 2022. Computer simulation modelling of the simultaneous adsorption of Cd, Cu and Cr from aqueous solutions by agricultural clay soil: A PHREEQC geochemical modelling code coupled to parameter estimation (PEST) study. Ecol. Model. 465, http://dx.doi.org/10.1016/j.ecolmodel.2022.109872.

Müller, M., D.L. Parkhurst, S.C., 2011. Programming PHREEQC calculations with C++ and Python - A comparative study. MODFLOW More 2011 - Integr. Hydrol. Model., Proc. 632–636.

Neal, C., Neal, M., Reynolds, B., Maberly, S.C., May, L., Ferrier, R.C., Smith, J., Parker, J.E., 2005. Silicon concentrations in UK surface waters. J. Hydrol. 304 (1–4), 75–93. http://dx.doi.org/10.1016/j.jhydrol.2004.07.023.

Neteler, M., Bowman, M., Landa, M., Metz, M., 2012. GRASS GIS: A multi-purpose Open Source GIS. Environ. Model. Softw. 31, 124–130. http://dx.doi.org/10.1016/j.envsoft.2011.11.014.

Owens, M., 2006. The Definitive Guide to SQLite. A Press.

Parkhurst, D., Appelo, C., 1999. US Geol. Surv. Water Resour. Inv. (99–4259), 597–643, URL https://www.usgs.gov/software/phreeqc-version-3.

Parkhurst, D.L., Appelo, C.A.J., 2013. Description of input and examples for PHREEQC version 3–A computer program for speciation, batch-reaction, one-dimensional transport, and inverse geochemical calculations. p. 497, URL http://pubs.usgs.gov/tm/06/a43.

Parkhurst, D., Kipp, K., Charlton, S., 2010. PHAST version 2—A program for simulating groundwater flow, solute transport, and multicomponent geochemical reactions. USGS. URL https://pubs.usgs.gov/tm/06A35/pdf/TM6-A35.pdf.

Pavuluri, S., Tournassat, C., Claret, F., Soulaine, C., 2022. Reactive transport modeling with a coupled OpenFOAM®-PHREEQC platform transport in porous media. Constr. Build. Mater. 145, 475–504. http://dx.doi.org/10.1007/s11242-022-01860-x.

Piper, A., 1945. A graphic procedure in geochemical interpretation of water analysis. Am. Geophys. Union Trans. 6914.

Pötter, L., Tollrian, R., Wisotzky, F., Weiss, L., 2021. Determining freshwater pCO2 based on geochemical calculation and modelling using PHREEQC. MethodsX 8, 101430. http://dx.doi.org/10.1016/j.zool.2021.125909.

Pourbaix, M., 1966. Atlas of Electrochemical Equilibria in Aqueous Solutions. Pergamon Press, Oxford.

PROJ contributors, 2022. PROJ coordinate transformation software library. http://dx.doi.org/10.5281/zenodo.5884394, Open Source Geospatial Foundation. URL https://proj.org/.

Smedley, P.L., Kinniburgh, D.G., 2002. A review of the source, behaviour and distribution of arsenic in natural waters. Appl. Geochem. 17 (5), 517–568.

Smedley, P.L., Kinniburgh, D.G., 2023. Uranium in natural waters and the environment: Distribution, speciation and impact. Appl. Geochem. 148, http://dx.doi.org/10.1016/j.apgeochem.2022.105534.

Steefel, C.I., Appelo, C.A.J., Arora, B., Jacques, D., Kalbacher, T., Kolditz, O., Lagneau, V., Lichtner, P.C., Mayer, K.U., Meeussen, J.C.L., Molins, S., Moulton, D., Shao, H., Šimůnek, J., Spycher, N., Yabusaki, S.B., Yeh, G.T., 2014. Reactive transport codes for subsurface environmental simulation. Comput. Geosci. 19, 445–478.

Stiff, Jr., H.A., 1951. The interpretation of chemical water analysis by means of patterns. J. Pet. Technol. 3 (10), 376–379.

The HDF Group, 2000–2010. Hierarchical data format version 5. URL http://www.hdfgroup.org/HDF5.

UFAM, 2023. National groundwater monitoring. URL https://www.bafu.admin.ch. (Accessed 13 January 2023).

USGS, 2016. National water information system. URL https://waterdata.usgs.gov/nwis/gw. (Accessed 13 January 2023).

van der Lee, J., Lomenech, C., 2004. Towards a common thermodynamic database for speciation models. Radiochim. Acta 92, 811–818.

Wissmeier, L., Barry, D., 2010. Implementation of variably saturated flow into PHREEQC for the simulation of biogeochemical reactions in the vadose zone. Environ. Model. Softw. 25, 526–538. http://dx.doi.org/10.1016/j.envsoft.2009.10.001.

Wolery, T.J., USDOE, 2010. EQ3/6 a software package for geochemical modeling. http://dx.doi.org/10.11578/dc.20210416.44, URL https://www.osti.gov//servlets/purl/1231666.

WRIS-India, 2023. WRIS India. URL https://indiawris.gov.in/wris/#/GWQuality. (Accessed 13 January 2023).