Environmental Sciences Europe

RESEARCH Open Access

SWAT + input data preparation in a scripted workflow: SWATprepR

Svajunas Plunge[1,2*], Brigitta Szabó[3], Michael Strauch[4], Natalja Čerkasova[5,6], Christoph Schürz[4] and Mikołaj Piniewski[1]

Abstract

Input data collection, quality assurance and preparation are central but time_consuming steps in environmental modeling. Errors due to manual processing of model input data can result in an incorrect representation of an environmental system and may consequently lead to implausible model simulations. Correct input data preparation and thorough quality check at an early stage of the model setup procedure are essential to build confidence in model simulation results. Typically, in environmental model applications, many steps in the input data preparation phase have to be repeated with the inflow of new, additional or corrected data. In this study, we selected the widely used SWAT + ecohydrological model as an illustrative example to investigate challenges related to input data preparation. To assist in these tasks, we developed an R package named SWATprepR, which provides functions for typical and repeating SWAT + model input data preparation tasks. The package supports the preparation of weather input files, atmospheric deposition, soil parameters, crop rotations, and observed (control or calibration) data, to name a few, presently with focus on European applications. The SWATprepR functions are integrated in R script workflows and can help SWAT + modelers to avoid repetitive tasks, secure reproducibility and transparently document the data processing steps. Application of the package is illustrated with a test case of a SWAT + model for a small catchment in central Poland.

Keywords SWAT + model, Input data processing, R package, Workflow, Reproducibility

Introduction

Rapid changes in the global environment bring challenges for the protection of ecosystems, which demand evidence-based policy making [1]. Environmental modeling is an essential part of it, as policy makers can better understand the potential impacts of their decisions on the environment and identify the most effective strategies for mitigating or adapting to environmental challenges [2]. Transparency of methodology and reproducibility are crucial prerequisites for modeling studies intended to inform policy decisions [3]. Surprisingly, even among published peer-reviewed studies, these essential elements are often absent [4, 5]. This deficiency is creating legal challenges for policymakers, where environmental decision-making is based on modeling as input data, parameters, model assumptions and validation processes are most often questioned [3]. Furthermore, missing

*Correspondence:
Svajunas Plunge
svajunas_plunge@sggw.edu.pl
[1] Department of Hydrology, Meteorology and Water Management, Warsaw University of Life Sciences, Nowoursynowska st. 159, 02-776 Warsaw, Poland
[2] Institute of Water Resources Engineering, Vytautas Magnus University, Universiteto st. 10, 53361 Akademija, Lithuania
[3] Institute for Soil Sciences, Centre for Agricultural Research, Herman Ottó út 15, Budapest 1022, Hungary
[4] Department Computational Landscape Ecology, Helmholtz Centre for Environmental Research GmbH—UFZ, Permoserstraße 15, 04318 Leipzig, Germany
[5] Marine Research Institute, Klaipėda University, Universiteto Ave. 17, 92294 Klaipėda, Lithuania
[6] Texas A&M AgriLife Research, Blackland Research and Extension Center, Temple, TX, USA

transparency and reproducibility in hydrological modeling studies raises concerns about the scientific quality of the results. This has resulted in an increasing demand from funding agencies and journals for the disclosure of the original data and code used in computations, highlighting its crucial role in scientific quality control [6].

The idea of scripted workflows designed for environmental modeling has been provided to solve this weakness in modeling studies [5, 7]. The main principle is that a common set of scripts is provided, with components designed to download and process input data, restructure model inputs to conform with required formats, run scenarios, extract results and compare them with the "baseline" or each other. If those scripts are prepared with commonly used open source scripting languages such as Python [8] or R [9] and released as packages via software sharing, collaboration and version control platforms (as GitHub, GitLab, Bitbucket, etc.), then multiple possibilities for collaboration on further development and tailored adaptation of the tools are presented. However, these tools have to be properly generalized and documented, which is rarely the case as modelers primarily use scripting to aid in their own modeling applications.

For example, the Soil and Water Assessment Tool (SWAT), a semi-distributed process-based ecohydrological modeling tool, has been employed for a duration exceeding a couple of decades [10]. It is free, open source and has been used worldwide for a great variety of surface water environment-related questions [11–14]. Records in the SWAT Literature Database provide around 6000 scientific papers [15]. Its official website also provides multiple solutions for model tailoring to different questions or tools to prepare different inputs. However, the application of those tools in sequential order requires adaptation to multiple file formats, software installations, understanding of different graphical user interfaces and involves a lot of manual manipulation, which is difficult to connect programmatically and is error-prone.

Some open source scripted tools have been made available for SWAT and the most recent SWAT + model versions [16]. As an example, SWAT + AW [5] presented a user-friendly, Python-based scripted workflow for catchment modeling. This tool utilizes preprocessed input data to facilitate the assemblance of a SWAT + model setup. Other Python-based examples to be mentioned are PySWAT [17], swatpy [18] or SpotSWATplus, which are mainly designed for coupling SWAT + /SWAT with the SPOTpy library [19] to edit model parameters and run it. However, those packages are not in active development at the moment. Currently (as of July 2022), two packages are promoted on the official SWAT model website[1] which are developed using R: SWATrunR [20] and R-SWAT [21]. These packages are designed primarily for model sensitivity assessment, calibration, validation and uncertainty analysis. Their application examples are reported in several studies [22–24] and, based on development records in GitHub repositories, these packages are actively developed and updated. There are other available SWAT + related open source R packages in active development, such as: SWATdoctR [25], designed for model setup verification; SWATfarmR [26]—a tool for preparing advanced agricultural management schedules for SWAT models, SWATbuildR [27, 28]—a comprehensive tool for building SWAT + model setups that includes connectivity between spatial objects. Those packages have been developed, used, updated, and tested within the EU-funded Horizon 2020 research and innovation OPTAIN project (OPtimal strategies to reTAIN and re-use water and nutrients in small agricultural catchments across different soil-climatic regions in Europe) and described in the project's modeling protocol [28].

Presented packages could be connected to a scripted workflow. Nevertheless, the tools mentioned necessitate preprocessed data, overlooking the labor-intensive process of preparing/preprocessing input data. Collection of such data, quality checks and preparation are probably the most time consuming stages in environmental modeling, which require proper diligence and verification to ensure smoothness of modeling effort in later steps [29]. Open source scripted tools provide automatization to save time and proper documentation, how raw data were treated. Moreover, corrected or updated data frequently become available during the lifetime of projects. Thus, having a scripted workflow covering the input data preparation step is highly beneficial.

Additionally, during an input data preparation stage, numerous questions could arise for the modeler. For instance, data have been obtained, but how accurate, complete, consistent is it? How to identify outliers and how to treat them correctly (remove or correct)? What data cover the temporal and spatial resolution needed? How to obtain important model input data or parameters, which are not measured in the field, and how to fill data gaps?

Another set of questions are connected to data format handling. For instance, which units are needed for a model and how to convert to them? What format and in which structure data should be delivered for a model? What files have to be provided or updated so the model finds and uses delivered data in the correct way? How to

[1] The link to website https://swat.tamu.edu/software/

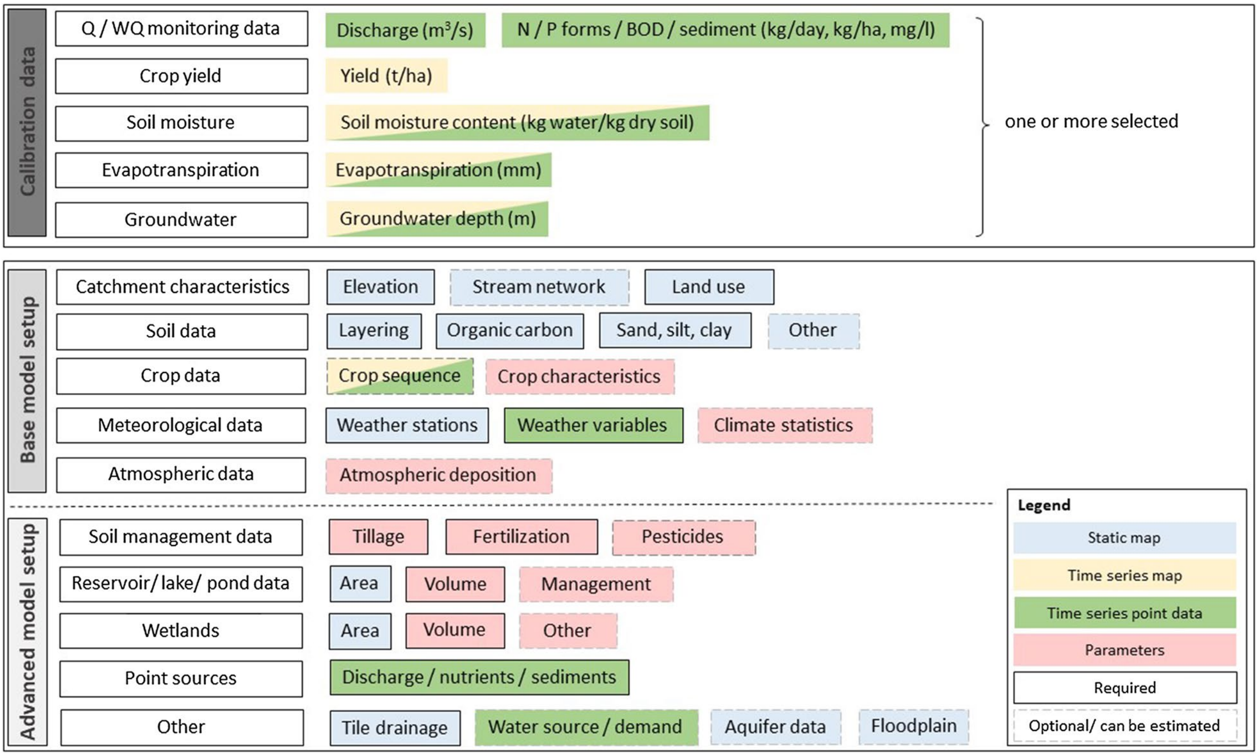

Fig. 1 Data requirements for SWAT/SWAT + model

track data preparation, handling and correct mistakes? How to document data handling so it could be reported and also used to add data in later stages, if more data become available or if someone else needs to update a modeling project, etc.? These are just a few examples of simple questions, which are likely to arise during the data preparation stage. Answering and providing solutions for them is time consuming. A valuable way to move forward and save modelers' time would be through a systematic approach that offers open-source tools with pre-existing answers to some of these questions, along with a flexible framework to seamlessly incorporate new solutions.

To further foster the use of automatized and transparent open source modeling workflows, this article introduces a new tool developed as an R package named SWATprepR with a collection of functions to preprocess input data and derive some missing parameters for the SWAT + model by providing demonstration examples of the tool's functionality for one case study. The current version of the tool covers the important steps of input data preparation in SWAT + modeling, including weather and climate, atmospheric deposition, soil parameters, crop rotation, observation, and point source data. Since the package is open-source, it allows users to easily integrate their own solutions to enhance its capabilities and address additional requirements.

SWATprepR package features

The SWAT + model is a process-based, semi-distributed, small watershed to river basin-scale model, and it requires multiple types of input data [30]. These data should be identified, collected, quality assessed, cleaned and transformed to model usable formats. Proper input data preparation is often the most labor intense and prolonged phase in the modeling process. Figure 1 provides an overview of the main data required by the model.

SWATprepR 1.0.2 version of the package includes functions which provide solutions to six different topics: weather data, atmospheric deposition, climate projection data, soil parameters, crop rotation and point source data. Functions can be categorized as designed for (i) loading data from the templates or online sources; (ii) plotting and cleaning data; (iii) calculating missing model parameters/data; and (iv) writing model input files (see Table 1). Detailed examples of application are presented in the https://biopsichas.github.io/SWATprepR/ website. This section is used to present an overview of existing functionalities. Yet other functionalities are easily accessible with different R packages as well (examples provided on package website).

Table 1 SWATprepR package features and functionality (X-functionality is supported in version 1.0.0)

Package feature	Functionality			
	Loading data (templates or online sources)	Plotting and cleaning data	Calculating missing parameters	Writing input files
Weather and climate input	X	X	X	X
Atmospheric deposition input	X			X
Soil parameters	X		X	X
Crop rotation data	X			
Observation data	X	X	X	
Point source data	X			X

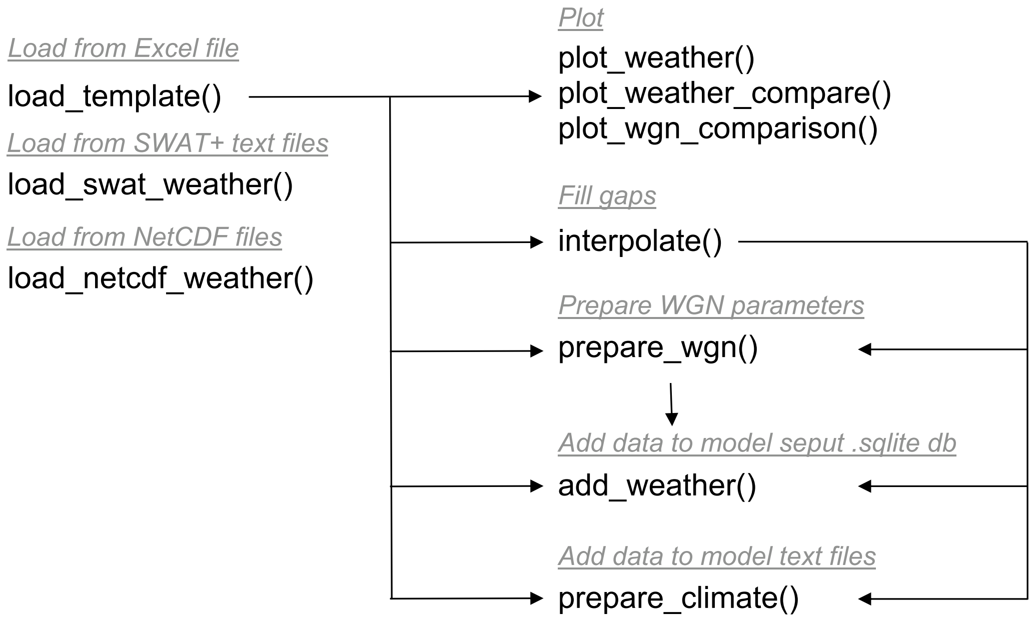

Fig. 2 Main functions and their functionalities for weather data in SWATprepR package

Weather and climate input

Weather data hold significant importance for hydrological models as it determines the form of precipitation (whether it's solid or liquid) and drives major water fluxes: e.g., evapotranspiration, and water flows within various media. Consequently, significant emphasis must be placed on quality assurance for meteorological variables. The SWATprepR package presents multiple options in addressing this concern, as depicted in Fig. 2.

The main function is load_template(), which loads data from the Excel[2] template (named '*weather_data.xlsx*') included in the package. The template requires typical information for meteorological stations, such as name, coordinates, altitude, and available time series for variables required by the SWAT + model (precipitation, temperature, wind speed, humidity, solar radiation). The function loads data into the R environment in a specific object, represented as a nested list format, which is used by this package. Once the data have been imported, the user can apply all other functions to the object. Two other functions could be applied to load weather data from different formats to the same object. The load_swat_weather() function can be used to load weather data directly from SWAT + input text files and the load_weather_netcdf() function can be applied to load weather data directly from Network Common Data Form (NetCDF) files format [31], which is often used to store large datasets, such as climate time series data.

Loaded data can be examined with multiple functions. For example, function plot_weather() can be used to

[2] Excel format was chosen as the most commonly used format for handling spreadsheet data.

perform a quality check on time series data. This function generates an interactive plot that shows data from all available stations. It offers various options for aggregation over multiple time intervals and provides different summarization functions such as mean, median, sum, standard deviation, minimum, maximum, and more. By using the plot_weather_compare() function, the user can extend this capability to compare two datasets. Function plot_wgn_comparison() generates a plot for comparing weather statistical values for two datasets, which might be needed in the assessment of weather data from projected climate datasets.

Upon loading and inspecting the data, the modeler may encounter situations in which certain stations have data gaps of different length. In such cases, different methods can be applied to fill in these gaps. For this purpose, the package provides the interpolate() function. To use this function, the modeler is required to provide a basin shapefile and a DEM raster file for the catchment area. Based on a user-defined grid spacing interval, the function creates virtual stations with interpolated weather variable data. The interpolation process is performed using the inverse distance weighting (IDW) method with a user-defined exponent parameter [32]. While there exist more sophisticated techniques for spatial interpolation of weather data [33], the IDW method has been widely used in different contexts for all weather variables required by SWAT + [34].

Another important input to SWAT + are weather statistical parameters, referred to as the input to the weather generator (WGEN or WGN). SWAT + uses the weather statistical data to fill gaps in daily weather data for short periods of time and to calculate plant growth initiation parameters. Despite the capability of in-build weather generator, the statistical data and its functionality are not recommended for simulating extended periods of missing data. To assist with this input, the official SWAT model website offers various tools, including the WGN Parameters Estimation Tool [35], WGN Excel macro [36], SWAT Precipitation Input Preprocessors and Dewpoint Estimation [37]. These tools require data preparation in different formats and demand familiarity with their respective functionalities. By using the SWATprepR package, in contrast, the modeler can calculate the required parameters with just a single command: prepare_wgn(), provided that weather data have been imported into the R environment.

The remaining two functions, add_weather() and prepare_climate(), provide two options to write weather data into the SWAT + model setup database. The first function requires three elements: an object containing loaded weather data, an object containing the calculated weather generator parameters, and the SWAT + model setup database in.sqlite format. The prerequisite is that the setup database should not have any pre-existing weather station or weather generator parameters entered into it. The add_weather() function adds weather time series, weather station data, and weather generator parameters to the model setup. The prepare_climate() function provides the option to transform the weather time series data directly into the formatted text files, which are used by the SWAT + model executable.

Atmospheric deposition input

SWAT + provides the option to include the observed atmospheric nitrogen deposition data into the model simulation. The input file requires a reduced (NH_4) and oxidized form (NO_3) of nitrogen in dry (kg/ha/year) and wet deposition (mg/l). These data may be available at specific locations or collected with field measurements. Another source of such data are atmospheric models [38]. SWATprepR supports the extraction of atmospheric deposition data from such models and adds it to the SWAT + model setup database. The function get_atmo_dep() uses the basin boundary shapefile as an input and downloads the required atmospheric deposition data directly from Meteorological Synthesizing Centre—West (MSC-W) model output data provided by the European Monitoring and Evaluation Programme (EMEP). The EMEP domain covers the geographic area between 30° N-82° N latitude and 30° W-90° E longitude [38].

Another function available in the current version of the package is add_atmo_dep(). To utilize it, the output from the get_atmo_dep() function and the path to the model setup database in.sqlite format is required. This function enables the incorporation of atmospheric deposition data into the model, allowing for the inclusion of deposition data ranging from daily to annual averages for a single station (Additional file 1).

Soil parameters

Soil characteristics are an important factor in determining the pathways of water when it reaches the land surface. Accurate soil parameters describing water and nutrient retention capacity and flow conditions in the soil matrix are essential for a successful and reliable SWAT + modeling study. Despite the current advancements in observations and data availability, in many study areas complete sets of required soil parameters are difficult, if at all possible, to obtain. Therefore the SWATprepR package includes a function, get_usersoil_table(), that simplifies the process of generating the complete set of soil parameters needed for the SWAT + model, i.e., moist bulk density, available water capacity, saturated hydraulic conductivity, moist albedo and Universal Soil Loss Equation soil erodibility factor. These parameters

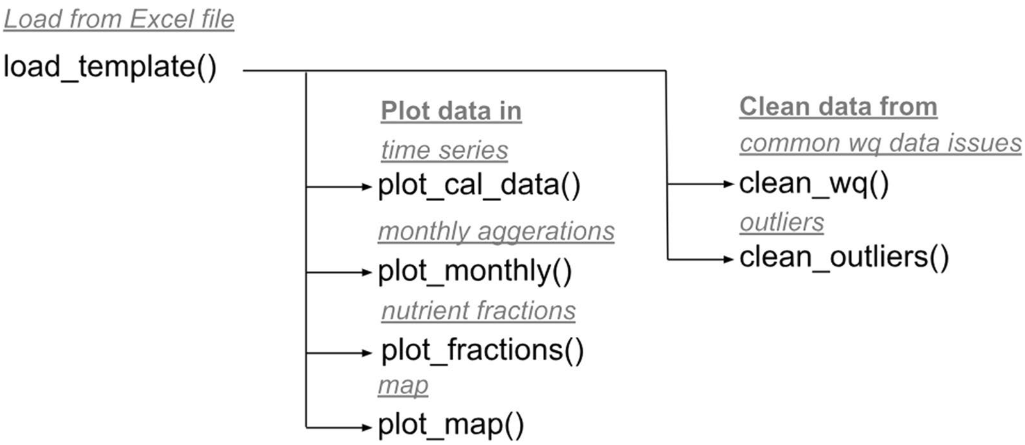

Fig. 3 Functions related to calibration/validation data in the SWATprepR package

are derived automatically from commonly available soil datasets by pedotransfer functions and other equations available from the literature [39–44]. To populate the SWAT + user soil parameter table, the modeler needs the following information for each soil layer within distinct soil types:

- Depth of layer;
- Percentage of clay defined as particles < 2 µm;
- Percentage of silt—2–50 µm;
- Percentage of sand—50–2000 µm;
- Soil organic carbon content in %.

For assigning Hydrologic Soil Groups, tile drainage, depth to groundwater level and impervious layer data are needed. The theoretical documentation for the pedotransfer functions integrated in this function is presented in the Sect. "Lacking crop rotation data" of the OPTAIN SWAT + modeling protocol [28].

Crop rotation data

The SWAT + model's popularity can be partly attributed to the capabilities to simulate the impacts of agricultural land management on water resources and water quality. To fully utilize this functionality, the modeler is required to supply information about agricultural activities representative of the study area, with one of the most crucial pieces being crop rotation data. Such information is rarely available or freely accessible, especially in large or transboundary watersheds. In such cases, remote sensing data could be utilized to generate information about crop rotations for the selected time period. Although the SWATprepR package itself does not have a function for directly extracting remote sensing data, it does offer functions to work with results of open source scripts that perform this task.

Google Earth Engine-based (GEE) remote sensing data extraction scripts were developed by Mészáros and Szabó [45] and described in a report of [46]. The script predicts crop types with a random forest method based on time series reflectance data of Sentinel 1A and 1B satellite radar images. The script generates a sequence of crop maps for each year as its output. To run this GEE script, the modeler needs the following input data: the shape of catchment boundary, continental or local crop data with coordinates as training points, and optionally, the boundaries of parcels or fields, if such data are available, which are added to the time series radar images selected based on the user-defined time period. The SWATprepR package provides two functions related to preprocessing (i) input data for the abovementioned GEE script and (ii) the derived time series crop map for the modeling. The first function, get_lu_points(), generates a set of training points for the remote sensing scripts. The second function, extract_rotation(), is designed for the extraction of crop rotation sequences per field. This data can be used with the SWATfarmR R package [26] to generate SWAT + model management input files.

Observation data

Observation data used for model calibration and validation are indispensable in most environmental model applications. Even though not strictly considered as model input data, they are required in the model preparation process for assessing and fine-tuning model performance. These data are usually collected and prepared along with other input data, and SWATprepR includes functions to quickly load, assess, plot, and clean monitoring data in the R environment (Fig. 3). It is important to emphasize that the SWATprepR functions were originally tailored for SWAT model users, but their versatility makes it straightforward to customize these functions for different variables or models as needed.

The load_template() function is used to load the data from the Excel template. A different template, named *calibration_data.xlsx*, is used to format the calibration data. After loading, the plot_cal_data() function can be applied to examine calibration and/or validation data time series for single or all the available gauge stations. Additionally, the plot_monthly() and plot_fractions() functions are designed for examining monthly aggregates and changes between ratios of different constituents. The plot_map() function is used to display the time series variables and station locations on interactive maps for assessing data availability, quality and variability in space.

Two basic functions are included in the package to aid with identifying and correcting errors and inconsistencies in the time series data. The clean_wq() function can fix most common issues related to water quality observation data, such as addressing comma-dot misuse, converting units by applying molecular weight conversion factors (e.g., converting to active substance weight as NH_4 to N-NH_4), handling negative values, removing missing values, and updating zero-concentrations to minimum positive values (or defined part of it), etc. Another function, clean_outliers(), allows the user to identify and remove data outliers in the time series. Outliers are identified as values outside the defined range (as mean ± n * standard deviation) of values.

Point source data

Point sources are generally considered to represent municipal or industrial wastewater treatment plants' discharge of treated sewage into the stream network. Discharge locations as well as the volume of effluents and the chemical characteristics of discharged water are typically needed to accurately represent the anthropogenic point source influence in a water quality model. The load_template() function in the SWATprepR package is used to load the data from an Excel template. An example template is included as a *pnt_data.xlsx* file. Once loaded into the R environment, the data can be examined for spatial and temporal consistency (using functions of ggplot2 or similar packages). The prepare_ps() function can be used to transfer the point source data into SWAT + model input file format. This function only requires a loaded point source object and a path to the model setup text files.

SWATprepR demonstration case

The developed SWATprepR tool was applied in a test case study of the Upper Zgłowiączka catchment. This catchment spans an area of 150 km^2 and is situated in central Poland. According to observation data for 2021, approximately 89% of this catchment is covered by arable land, while pastures account for 2%, forests for 5.5%, and urban and water areas for 3.5%. About 59% of the catchment's territory is equipped with tile drains. The case study site featured two point sources, 14 meteorological stations situated both within and around the catchment, as well as 21 water quality and flow measurement stations. This specific case study site was selected as one of the 14 sites within the OPTAIN project [47]. Within the scope of the project, a fine scale SWAT + model is set up and used to facilitate the evaluation of environmental effectiveness of Natural/Small Water Retention Measures (NSWRMs). The detailed setup of the SWAT + model necessitated the collection of various types of data, aimed at providing comprehensive environmental insights into the local conditions. Figure 4 illustrates a subset of the geospatial data that were gathered for the selected catchment. The process of collecting this detailed data and preparing the models input data posed several challenges, all of which were successfully addressed through the utilization of functions within the SWATprepR package. Below, we provide several illustrative examples.

Scarce weather data

Only one meteorological station was located inside the catchment. Yet it had data only for around 8 years, which was not enough for the foreseen modeling purposes. Additional meteorological data were collected from 13 stations in the vicinity of the catchment (within 40 km radius). All the collected meteorological data have been loaded with the SWATprepR load_template() function. Following that, the interpolate() function was applied to create a series of virtual weather stations within the catchment and the interpolation process was conducted for each day throughout the time series. This approach is a fast way to prepare a consistent spatially distributed meteorological data set for the catchment. Additionally, the actual meteorological stations had multiple gaps in the observation time series, which could be addressed with the interpolation procedure. Figure 5 provides an average percentage of available time series data within the catchment for all required variables for each meteorological station within the period of 1998–2022. By using the SWATprepR package, we generated a 2-km spaced grid that resulted in 38 virtual stations, which had a 100% data coverage for the selected time period and were located in the catchment.

Insufficient soil parameter information

Collecting soil parameters necessary for the SWAT + model at a detailed level can pose challenges. For the Upper Zgłowiączka catchment soil type map with values of sand, silt, clay, and soil organic carbon content

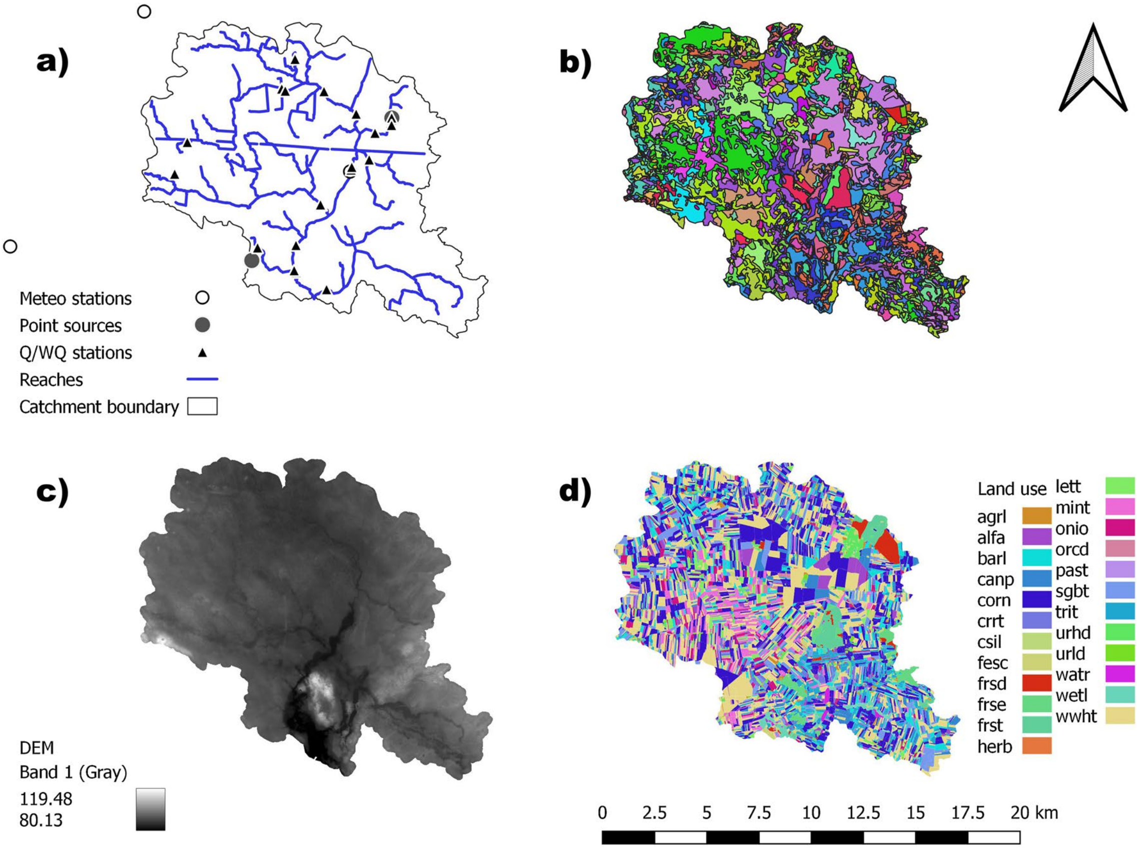

Fig. 4 Upper Zgłowiączka GIS catchment data. **a** Water flow, water quality (Q/WQ), meteorological stations, point source locations, reaches and catchment boundary; **b** soil type map; **c** DEM map; **d** land use map with crop type specification for 2021, classes defined as in land cover/plant growth database [48]

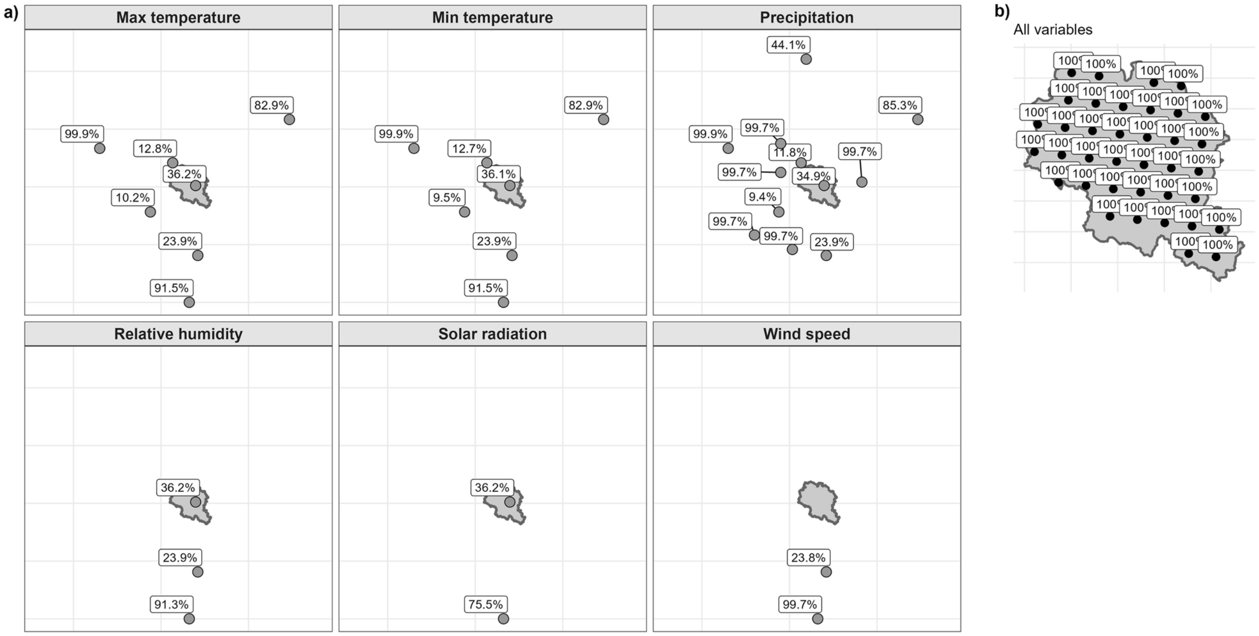

Fig. 5 Meteorological stations selected for data collection with the evaluation of data coverage in percentage for each variable (**a**) and virtual stations created with data coverage in percentage for all variables (**b**)

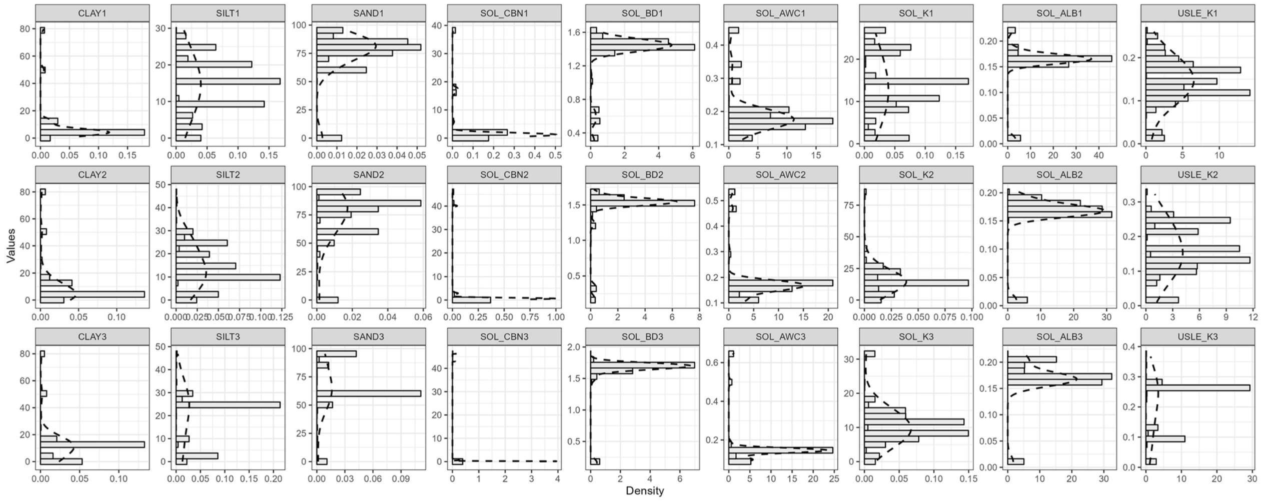

Fig. 6 Distribution of soil parameters values across three soil layers. Clay, silt, sand and soil organic carbon content (SOL_CBN) are used as get_usersoil_table() function input. Moist bulk density (SOL_BD), available water capacity (SOL_AWC), saturated hydraulic conductivity (SOL_K), moist soil albedo (SOL_ALB) and Universal Soil Loss Equation soil erodibility factor (USLE_K) calculated by the get_usersoil_table() function

were available for each soil type with characteristic soil layering. The availability of basic soil information allowed the parameterization of required SWAT + soil parameters by the get_usersoil_table() function. This function utilizes sand, silt, clay, and carbon content to parameterize moist bulk density, available water capacity, saturated hydraulic conductivity, moist soil albedo, and the Universal Soil Loss Equation soil erodibility factor. Figures 6 and 7 show the derived soil parameters of the function for the studied catchment. Additionally, hydrologic soil groups could be computed based on the data available on tile drainage, groundwater depth, and depth to water-impermeable layer. Detailed description of pedotransfer functions and methodologies applied to the calculation of parameters is presented in the SWAT + modeling protocol pages 81–92 [28].

Unavailable atmospheric deposition data

No locally collected atmospheric deposition data were available for the Upper Zgłowiączka catchment. SWATprepR the get_atmo_dep() function was used to retrieve atmospheric deposition data for the catchment and another package's function the add_atmo_dep() was used for incorporating downloaded data into the SWAT + model setup. Figure 8 illustrates an example of atmospheric deposition data extracted and plotted for the case study area, where previously no data was available.

Lacking crop rotation data

In the case of the Upper Zgłowiączka catchment, crop data were accessible only for a single year—2021. This proved insufficient for generating the required crop rotation sequences for the foreseen modeling task. To address this issue, a Google Earth Engine-based script [45] was applied to identify crops for field parcels for previous years using Sentinel 1A and 1B satellite radar images. For the training of the crop classification model, local crop data were used. The crop maps were generated with a tool developed and validated as a part of the OPTAIN project [46]. Next, the SWATprepR extract_rotation() function was applied and field-based annual crop sequences were extracted. The obtained results are presented in Fig. 9. The catchment is predominantly characterized by winter wheat cultivation, accounting for approximately 20% of all rotations, while winter wheat with corn represents another 13%, and winter wheat with sugar beets accounts for 7%. Additionally, corn-to-corn rotations make up 6%, winter wheat to barley—5.5%. These outputs were then utilized in conjunction with the SWATfarmR tool [26] to develop crop management schedules tailored to the specifics of the SWAT + model.

Limitations

There are several limitations of the SWATprepR package that users should be aware of. It is challenging to list all potential limitations due to the diverse application cases users might have in mind. Nonetheless, we would like to highlight some examples to provide users with a better understanding of the current shortcomings in the SWATprepR version.

The current version (as of January 2024) of interpolate() function only incorporates the IDW interpolation technique for weather data. Its effectiveness depends on factors such as location, topography, variable, and

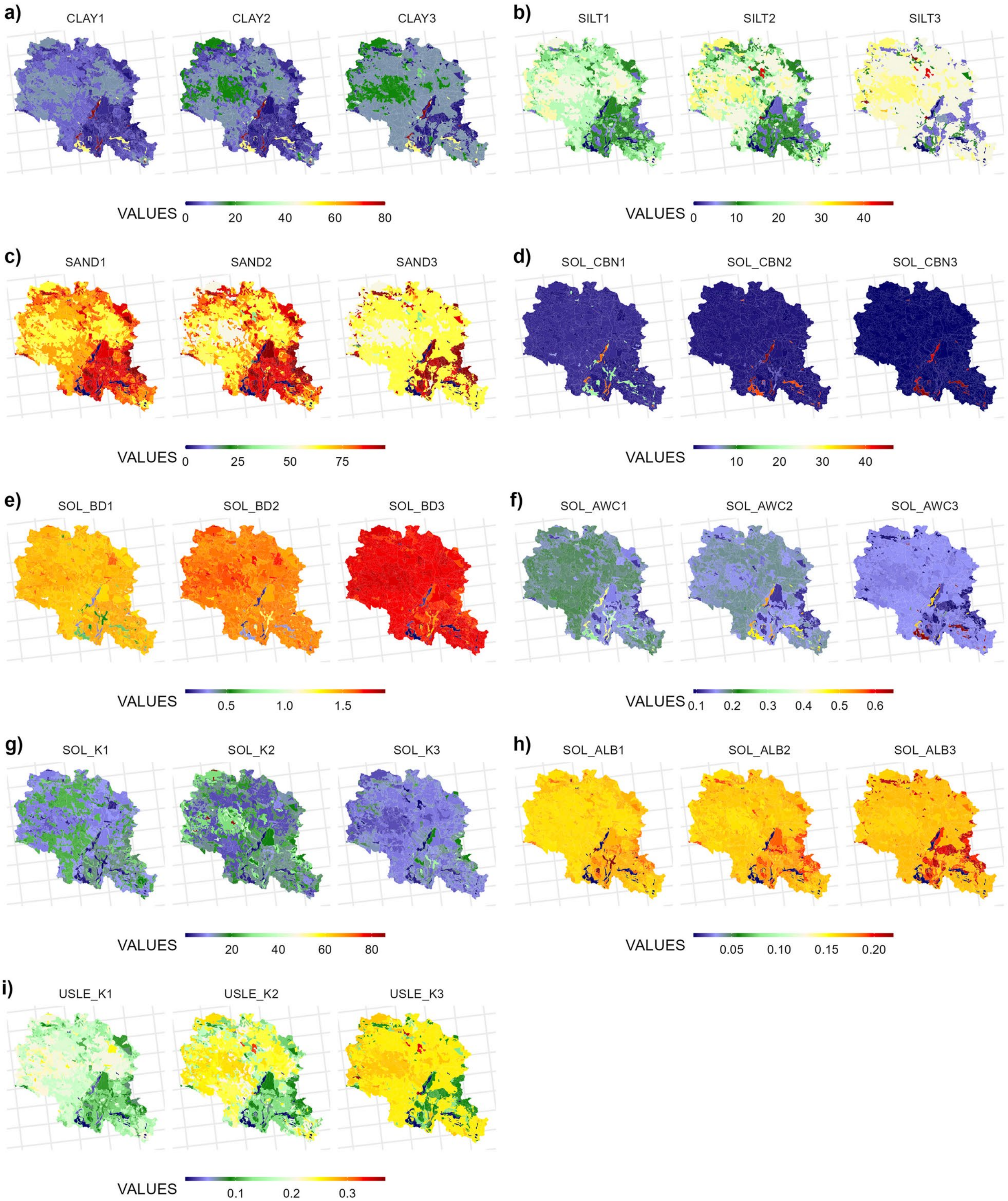

Fig. 7 Maps of calculated soil parameters for all three soil layers: **a** clay, **b** silt, **c** sand, and **d** soil organic carbon content, **e** moist bulk density, **f** available water capacity, **g** saturated hydraulic conductivity, **h** moist soil albedo, **i** Universal Soil Loss Equation (USLE) soil erodibility factor

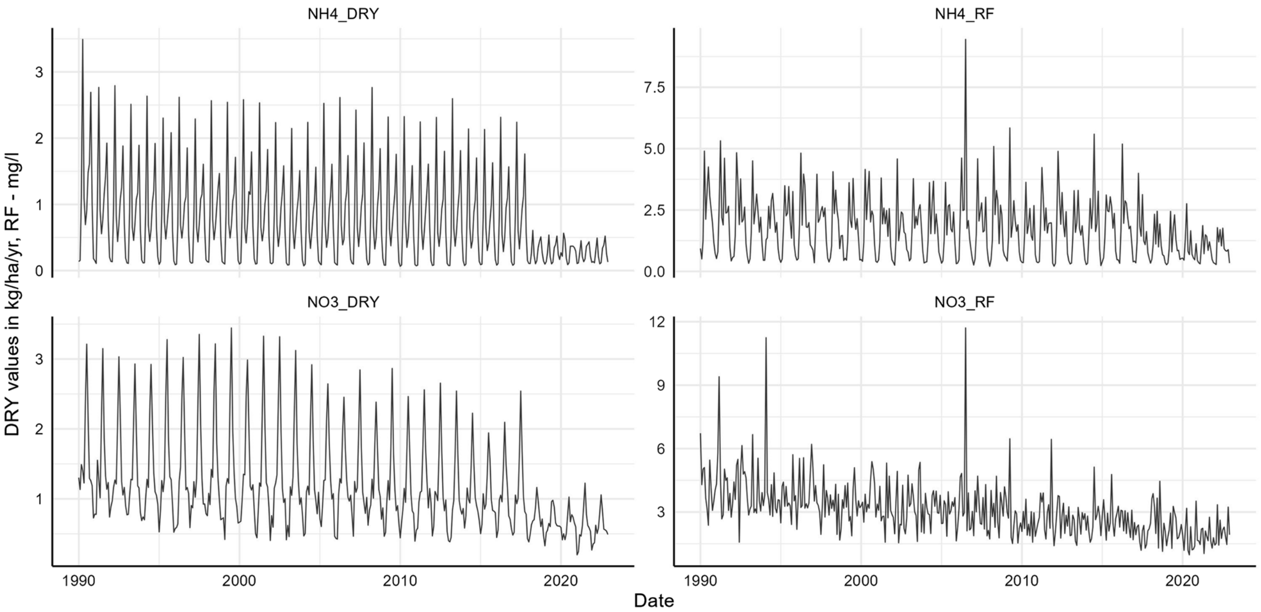

Fig. 8 Atmospheric deposition data for the Upper Zgłowiączka catchment. *NH4_DRY* ammonia dry deposition (kg/ha/year), *NH4_RF* ammonia in rainfall (mg/l), *NO3_DRY* nitrate dry deposition (kg/ha/year), *NO3_RF* nitrate in rainfall (mg/l), *NH4_DRY* ammonia dry deposition (kg/ha/year), *NO3_DRY* nitrate dry deposition (kg/ha/year)

data gap length. The IDW method is widely used in meteorology because it is fast and easy to implement. However, ancillary data, such as elevation, cannot be incorporated and the method tends to generate “bull’s eye patterns” [49]. Moreover, as there is no extrapolation, all interpolated values are within the range of the data points [50]. Customization for alternative (e.g., probabilistic) methods, such as kriging [49], may be necessary, as IDW interpolation might not be suitable in certain conditions.

The atmospheric deposition function get_atmo_dep() prepares inputs for the EMEP data domain for Europe, parts of northern Africa, and western Asia. It is not applicable to regions outside of this domain. The generated data are based on outputs of the MSC-W model [51] and thus afflicted with uncertainties. Additionally, EMEP updates its calculations yearly, providing information under new server links with slightly different coding, making it impossible to obtain the latest data without adjusting the function. Therefore, the get_atmo_dep() function is tailored to the last available version of EMEP data at the time of article preparation. Users should be aware of this if they intend to use the latest EMEP data, and tailor the functionality according to their needs. Users should also mind that the current version of the get_atmo_dep() function generates basin-averaged single time series without distinguishing between regions or stations, which may be an important shortcoming for large-scale model applications.

The organic carbon content for each soil layer, required for the get_usersoil_table() function, can be challenging to obtain in many regions. Information needed for preparing soil hydrologic groups, such as impervious layer depth, depth to the high water table, or drainage status of soils, could be even more challenging to collect. Proxy data may be used, introducing potential inaccuracies and uncertainties. Based on soil texture and organic carbon content, the get_usersoil_table() function predicts hydrologically effective soil parameters, such as available water capacity and saturated hydraulic conductivity. The root mean squared error of the built-in pedotransfer functions were 0.048 cm^3 cm^{-3} for available water capacity and 1.48 cm day^{-1} for logarithmic ten transformed saturated hydraulic conductivity [52] on the test sets of the European Hydropedological Data Inventory point dataset [53]. This dataset includes temperate soils, the uncertainty of the pedotransfer functions in other regions is therefore unknown.

Functions related to crop rotations are linked to application of the GEE-based remote sensing script [45] and SWATfarmR [26], potentially limiting utility for users with other types of land use maps, who will not make use of the SWATfarmR functionality. In case the GEE-based crop classification is utilized, users should be aware of possible classification errors, which can be reduced by incorporating a sufficient amount of local training data.

Other technical limitations include interactive functions connected to plotting observation data, reliant on

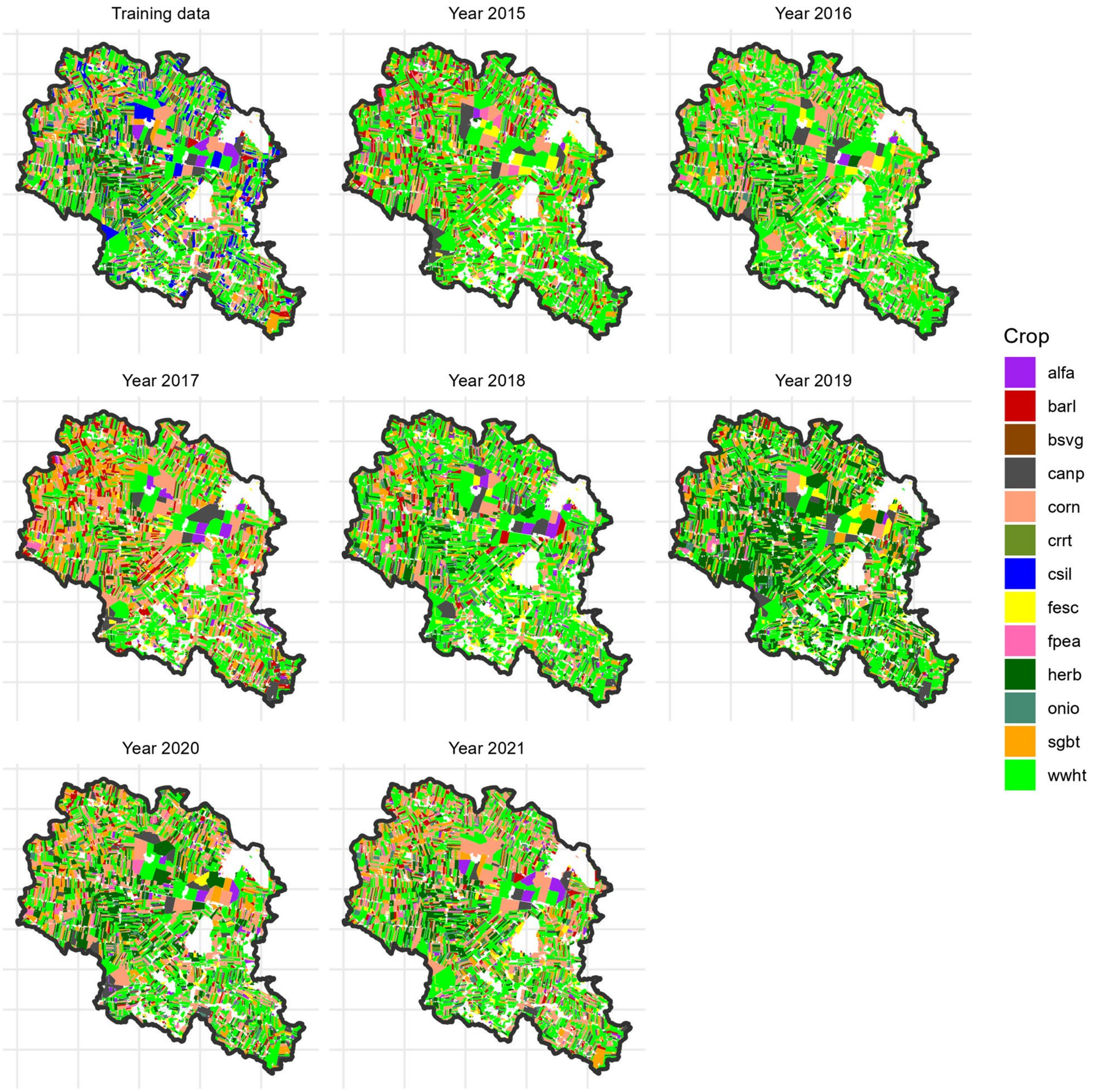

Fig. 9 Generated crop maps for each year and each field in the catchment. Training data are based on farmers' declarations for the year 2021. Meaning of crop codes available in SWAT model databases documentation available on https://swat.tamu.edu/media/69419/Appendix-A.pdf

the R plotly package [51], which may encounter issues displaying data of large datasets. The current SWATprepR version's point source data preparation function, prepare_ps(), does not load data with multi-annual averages, potentially posing challenges when constant point source loads are required in the model. The functions for data cleaning, clean_wq() or clear_outliers(), only include the most elementary techniques, without providing options for more advanced data cleaning methods, especially with regard to outlier detection [54, 55].

These examples do not cover all limitations of the SWATprepR package, as there are many possible use cases. Yet, they provide users with an understanding of potential methodological, data-related, or technical constraints they might encounter.

Conclusions and future work

The SWATprepR package offers valuable tools and incorporates effective techniques to assist SWAT + modelers in preparing their models. One of the primary challenges in creating a comprehensive model setup lies in the sheer volume of high-resolution spatial and temporal data required to feed into the model. With numerous variables, parameters and processes to adjust, modelers often find themselves overwhelmed, especially when data availability is not straightforward.

Another challenge is the number of different file formats (spreadsheets, text files, relational databases, NetCDF, etc.), which all require different tools or approaches to manipulate. Errors or biases in the input data may make the entire available datasets untrustworthy or unusable. Consequently, modelers may accidently introduce errors into their setup or choose to omit critical information, significantly diminishing model reliability. Unfortunately, such shortcomings are often masked through subsequent parameterization of processes in calibration, leading to unreliable simulations that can have far-reaching implications for decision-making by end-users of the model.

The use of scripted workflows like SWATprepR offers significant advantages. One of them is quick and easy error correction. When errors or inaccuracies are identified in the setup of SWAT models, scripted workflows make it straightforward to implement corrections. Instead of manually retracing and repeating steps, modelers can easily modify and rerun the scripts to ensure that the model setup aligns with the desired specifications. This saves time and reduces the likelihood of human errors during the correction process.

Another advantage is easy adaptation to new, updated datasets. As new or updated datasets become available or as the project requirements change, scripted workflows prove invaluable. Modelers can efficiently integrate new data sources into the existing model setup by updating scripts.

Furthermore, scripted workflows could also facilitate collaboration among modelers. When workflows are documented and shared, other researchers or modelers can readily understand the processes and parameters used in the SWAT modeling. This ease of comprehension allows for efficient collaboration, peer review, and the potential for others to build upon or extend the existing models. In addition, scripted workflows can be managed with version control systems like Git, ensuring a history of changes, easy tracking of modifications, and the ability to revert to previous states, if needed. This enhances the reproducibility and traceability of the modeling process. As well as automation through scripting ensures a high level of consistency across different runs of the SWAT model. This consistency is essential for producing reliable and comparable results in scientific research or environmental assessments.

The model setup insights and parameter estimation methods discussed in this paper should prove invaluable to any modeler embarking on a journey to establish a dependable case study analysis using SWAT +. The SWATprepR package is open-source and will continue to undergo active development and enhancement in the foreseeable future, reducing its current limitations mentioned in the previous section. We extend an invitation to the modeling community to contribute to the evolution of these tools, adapt our proposed methods to strengthen their own models, perform rigorous quality checks, and ultimately, contribute to a more transparent and informed decision-making process through the utilization of the SWAT + model.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12302-024-00873-1.

Additional file 1: SWATprepR package manual.

Software availability

Name of the software: SWATprepR. Developer: Svajunas Plunge. Contact information: svajunas_plunge@sggw.edu.pl. Year first available: 2022. Program language: R. Cost: free. Software availability: https://biopsichas.github.io/SWATprepR/. Program size: 163 kB.

Author contributions

SP software, formal analysis, visualization, writing—original draft, writing—review and editing. BS software, writing—original draft, writing—review and editing, methodology. MS software, writing—original draft, writing—review and editing. NČ writing—original draft, writing—review and editing, methodology. CS conceptualization, methodology. MP writing—original draft, writing—review and editing.

Funding

This work was accomplished within the OPTAIN project (OPtimal strategies to reTAIN and re-use water and nutrients in small agricultural catchments across different soil-climatic regions in Europe, cordis.europa.eu) which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 862756.

Availability of data and materials

The SWATprepR package, its documentation, and test data are freely available on https://github.com/biopsichas/SWATprepR.

Declarations

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 3 October 2023 Accepted: 20 February 2024
Published online: 11 March 2024

References

1. Pullin AS, Knight TM (2009) Doing more good than harm—building an evidence-base for conservation and environmental management. Biol Conserv 142:931–934. https://doi.org/10.1016/j.biocon.2009.01.010
2. Schmolke A, Thorbek P, DeAngelis DL, Grimm V (2010) Ecological models supporting environmental decision making: a strategy for the future. Trends Ecol Evol 25:479–486. https://doi.org/10.1016/j.tree.2010.05.001
3. Özkundakci D, Wallace P, Jones HFE et al (2018) Building a reliable evidence base: legal challenges in environmental decision-making call for a more rigorous adoption of best practices in environmental modelling. Environ Sci Policy 88:52–62. https://doi.org/10.1016/j.envsci.2018.06.018
4. Vos MG de, Janssen SJC, Bussel LGJ van, et al (2011) Are environmental models transparent and reproducible enough? MODSIM2011. In: 19th International Congress on Modelling and Simulation. https://doi.org/10.36334/modsim.2011.g7.devos
5. Chawanda CJ, George C, Thiery W et al (2020) User-friendly workflows for catchment modelling: towards reproducible SWAT+ model studies. Environ Modell Softw 134:104812. https://doi.org/10.1016/j.envsoft.2020.104812
6. Hutton C, Wagener T, Freer J et al (2016) Most computational hydrology is not reproducible, so is it really science? Water Resour Res 52:7548–7555. https://doi.org/10.1002/2016wr019285
7. Coon ET, Shuai P (2022) Watershed workflow: a toolset for parameterizing data-intensive, integrated hydrologic models. Environ Modell Softw 157:105502. https://doi.org/10.1016/j.envsoft.2022.105502
8. Python Software Foundation (2023) Python Language Reference, version 3.11. http://www.python.org. Accessed 10 Jan 2023
9. R Foundation (2023) The R Project for Statistical Computing 4.2. https://www.r-project.org/. Accessed 10 Jan 2023
10. Arnold JG, Srinivasan R, Muttiah RS, Williams JR (1998) Large area hydrologic modeling and assessment part I: model development. JAWRA J Am Water Resour Assoc 34:73–89. https://doi.org/10.1111/j.1752-1688.1998.tb05961.x
11. Tan ML, Gassman P, Yang X, Haywood J (2020) A review of SWAT applications, performance and future needs for simulation of hydro-climatic extremes. Adv Water Resour 143:103662. https://doi.org/10.1016/j.advwatres.2020.103662
12. Gassman PW, Reyes MR, Green CH, Arnold JG (2007) The soil and water assessment tool: historical development, applications, and future research directions. Trans ASABE 50:1211–1250
13. Gassman PW, Yingkuan W (2015) IJABE SWAT special issue: innovative modeling solutions for water resource problems. Int J Agric Biol Eng 8:1–8
14. Akoko G, Le TH, Gomi T, Kato T (2021) A review of SWAT model application in Africa. Water-sui 13:1313. https://doi.org/10.3390/w13091313
15. CARD&ISU (2023) SWAT literature database for peer-reviewed journal articles. https://www.card.iastate.edu/swat_articles/. Accessed 10 Jan 2023
16. Bieger K, Arnold JG, Rathjens H et al (2017) Introduction to SWAT+, a completely restructured version of the soil and water assessment tool. JAWRA J Am Water Resour Assoc 53:115–130. https://doi.org/10.1111/1752-1688.12482
17. Ferreira DB (2019) PySWAT: a Python application for Input/Output analysis for the Soil and Water Assessment Tool (SWAT). https://github.com/davidbispo/PySWAT. Accessed 10 Jan 2023
18. Kmoch A (2022) swatpy: A set of python modules to work with SWAT2012 models. https://doi.org/10.5281/zenodo.6322023
19. Houska T, Kraft P, Chamorro-Chavez A, Breuer L (2015) SPOTting model parameters using a ready-made python package. PLoS ONE 10:e0145180. https://doi.org/10.1371/journal.pone.0145180
20. Schürz C (2019) SWATrunR: running SWAT2012 and SWAT+ Projects in R. https://doi.org/10.5281/zenodo.3373859
21. Nguyen TV, Dietrich J, Dang TD et al (2022) An interactive graphical interface tool for parameter calibration, sensitivity analysis, uncertainty analysis, and visualization for the Soil and Water Assessment Tool. Environ Modell Softw 156:105497. https://doi.org/10.1016/j.envsoft.2022.105497
22. Musyoka FK, Strauss P, Zhao G et al (2021) Multi-step calibration approach for SWAT model using soil moisture and crop yields in a small agricultural catchment. Water-sui 13:2238. https://doi.org/10.3390/w13162238
23. Maref N, Baahmed D, Bemmoussat K, Mahfoud Z (2022) SWAT model application for sediment yield modeling and parameters analysis in Wadi K'sob (Northeast of Algeria). https://doi.org/10.21203/rs.3.rs-2069353/v1
24. Yang C, Xu M, Fu C, et al (2022) Glacier hydrological process modeling based on improved SWAT+: a case study in the Upper Yarkant River Basin. https://doi.org/10.22541/au.164512280.00856493/v1
25. Plunge S, Schürz C, Čerkasova N et al (2023) SWAT+ model setup verification tool: SWATdoctR. Environ Model Softw 171:105878. https://doi.org/10.1016/j.envsoft.2023.105878
26. Schürz C (2023) SWATfarmR: Simple rule based scheduling of management operations in SWAT. https://github.com/chrisschuerz/SWATfarmR. Accessed 10 Jan 2023
27. Schürz C (2022) SWATbuildR. https://git.ufz.de/optain/wp4-integrated-assessment/swat/bildr_script. Accessed 10 Jan 2023
28. Schürz C, Čerkasova N, Farkas C et al (2022) SWAT+ modeling protocol for the assessment of water and nutrient retention measures in small agricultural catchments. Zenodo. https://doi.org/10.5281/zenodo.7463395
29. ASABE N-21 H committee of (2017) Guidelines for calibrating, validating, and evaluating hydrologic and water quality (H/WQ) models. ASABE
30. SWAT+ Website (2023) SWAT+ Documentation. https://swatplus.gitbook.io/io-docs/. Accessed 27 Dec 2023
31. Unidata (2023) Network Common Data Form (NetCDF). https://www.unidata.ucar.edu/software/netcdf/. Accessed 29 Dec 2023
32. Shepard D (1968) A two-dimensional interpolation function for irregularly-spaced data. In: Proc 1968 23rd ACM Natl Conf. pp 517–524. https://doi.org/10.1145/800186.810616
33. Ma YZ (2019) Geostatistical estimation methods: kriging. In: Ma YZ (ed) Quantitative geosciences: data analytics, geostatistics, reservoir characterization and modeling. Springer International Publishing, Cham, pp 373–401
34. Ozelkan E, Bagis S, Ozelkan EC et al (2015) Spatial interpolation of climatic variables using land surface temperature and modified inverse distance weighting. Int J Remote Sens 36:1000–1025. https://doi.org/10.1080/01431161.2015.1007248
35. Essenfelder AH (2016) SWAT weather database: a quick guide. https://doi.org/10.13140/rg.2.1.4329.1927
36. Boisrame G (2011) WGNmaker4.xlsm manual. https://swat.tamu.edu/media/41586/wgen-excel.pdf. Accessed 10 Jan 2023
37. Liersch S (2003) The Programs dew.exe and dew02.exe User's Manual. https://swat.tamu.edu/media/83105/dewpoint.zip. Accessed 20 Sep 2023
38. MSC-W, CCC, CEIP, CIAM (2022) Transboundary particulate matter, photo-oxidants, acidifying and eutrophying components. https://emep.int/publ/reports/2022/EMEP_Status_Report_1_2022.pdf. Accessed 31 July 2023
39. Alexander EB (1980) Bulk densities of California soils in relation to other soil properties. Soil Sci Soc Am J 44:689–692. https://doi.org/10.2136/sssaj1980.03615995004400040005x
40. Assouline S, Or D (2014) The concept of field capacity revisited: defining intrinsic static and dynamic criteria for soil internal drainage dynamics. Water Resour Res 50:4787–4802. https://doi.org/10.1002/2014wr015475
41. Gascoin S, Ducharne A, Ribstein P et al (2009) Sensitivity of bare soil albedo to surface soil moisture on the moraine of the Zongo glacier (Bolivia). Geophys Res Lett 36:L02405. https://doi.org/10.1029/2008gl036377
42. Sharpley AN, Williams JR (1990) EPIC—erosion/productivity impact calculator: 1. Model documentation
43. Szabó B, Weynants M, Weber TKD (2020) Updated European hydraulic pedotransfer functions with communicated uncertainties in the predicted variables (euptfv2). Geosci Model Dev 14:151–175. https://doi.org/10.5194/gmd-14-151-2021
44. Wessolek G (2009) Bodenphysikalische Kennwerte und Berechnungsverfahren für die Praxis
45. Mészáros J, Szabó B (2022) Script to derive and apply crop classification based on Sentinel 1 satellite radar images in Google Earth Engine platform. https://doi.org/10.5281/zenodo.6700122

46. Szabó B, Mészáros J, Kassai P, et al (2022) Solutions to overcome data scarcity. Deliverable D3.2 EU Horizon 2020 OPTAIN Project, Grant agreement No. 862756
47. OPTAIN (2023) Upper Zglowiaczka. https://www.optain.eu/case-studies-and-actors-involvement/upper-zglowiaczka. Accessed 21 Sep 2023
48. Arnold JG, Kiniry JR, Srinivasan R, et al (2012) Soil & water assessment tool input/output documentation version 2012
49. Sluiter R (2009) Interpolation methods for climate data—literature review. KNMI
50. De Smith MJ, Goodchild MF, Longley P (2018) Geospatial analysis: a comprehensive guide to principles, techniques and software tools, 6th edn. Troubador publishing Ltd., Market Harborough
51. Simpson D, Benedictow A, Berge H et al (2012) The EMEP MSC-W chemical transport model–technical description. Atmos Chem Phys 12:7825–7865. https://doi.org/10.5194/acp-12-7825-2012
52. Szabó B, Kassai P, Plunge S, et al (2024) Addressing soil data needs and data-gaps in catchment scale environmental modelling: the European perspective. Manuscript submitted for publication
53. Weynants M, Montanarella L, Toth G, et al (2013) European hydropedological data inventory (EU-HYDI). https://doi.org/10.2788/5936
54. Thériault R, Ben-Shachar MS, Patil I et al (2023) Check your outliers! An introduction to identifying statistical outliers in R with easystats. PsyArXiv. https://doi.org/10.31234/osf.io/bu6nt
55. Jamshidi EJ, Yusup Y, Kayode JS, Kamaruddin MA (2022) Detecting outliers in a univariate time series dataset using unsupervised combined statistical methods: a case study on surface water temperature. Eco Inform 69:101672. https://doi.org/10.1016/j.ecoinf.2022.101672

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.