



**ADVERTIMENT.** L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons:  <https://creativecommons.org/licenses/?lang=ca>

**ADVERTENCIA.** El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons:  <https://creativecommons.org/licenses/?lang=es>

**WARNING.** The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  <https://creativecommons.org/licenses/?lang=en>

PhD Thesis

NOVEL ALGORITHMS FOR GEOLOGICAL  
AND HYDROGEOLOGICAL ASSESSMENTS

Ashkan Hassanzadeh

PhD Thesis

**NOVEL ALGORITHMS FOR GEOLOGICAL AND  
HYDROGEOLOGICAL ASSESSMENTS**

Thesis presented by:

**Ashkan Hassanzadeh**

Barcelona May 2023

Work conducted at:

Institute of Environmental Assessment and Water Research (IDAEA-CSIC)

Autonomous University of Barcelona (UAB)

Advisors:

Enric Vázquez-Suñé

Mercè Corbella





## Acknowledgments

Upon completing this challenge, I would like to convey my deepest appreciation for all the individuals who accompanied me on this journey, both directly and indirectly, and whose contributions made my PhD attainable.

Foremost, I am exceedingly grateful for my supervisors, Enric and Mercè, for their invaluable counsel, unwavering support, and patience throughout my doctoral studies. Their immense knowledge and abundant experience have been a source of encouragement to me throughout my academic pursuits.

I would like to acknowledge the colleagues at IDAEA, particularly the geoscience group on the fourth floor, for their warm and supportive attitudes, including Sonia, Ignacio, Franco, Rotman, Laura, and others whom I regret not mentioning by name.

I extend my heartfelt gratitude to my partner, Gemma, who stood by my side through the ups and downs of this period. Thank you for your unwavering support.

I would also like to recognize the bond of friendship that has been one of the most fulfilling, and valuable accomplishments of my life thus far. Meysam, Mohsen, Mahdi and other friends from my hometown have always been there for me, even though we live in different corners of this planet.

I would like to thank my family for laying the foundation of who I am today.

Lastly, I want to express my appreciation to all individuals who respect all forms of life to the best of their ability as human beings and who devote their lives to restoring this planet with selflessness

## Abstract

Groundwater numerical modeling is essential for sustainable water management as it helps understand the behavior of groundwater resources and their interaction with surface water. These models can predict the impacts of water manipulation on the environment, aid in developing sustainable management scenarios, and help in understanding the impacts of climate change on groundwater resources. Generating numerical groundwater models require various input data and boundary conditions, including geological models, surface water input, and groundwater recharge. By combining different modeling approaches, a comprehensive understanding of groundwater resources can be developed to support sustainable groundwater management, decision-making, and water resource planning. This thesis is based on open-source tools to aid in the different stages of groundwater studies: geological 3D modeling (Geopropy), isotopic modeling (Isocompy), and soil water balance spatial-temporal modeling (WaterpyBal and WaterpyBal Studio). Geopropy is a decision-making algorithm implemented in Python, generates 3D cross-sections. It performs as an intelligent agent that simulates the steps taken by the geologist in the process of creating the cross-section, coupled with data-driven decisions. The algorithm detects zones with more than one possible outcome and, based on the level of complexity (or user preference), proceeds to automatic, semiautomatic or manual stages. Isocompy is Python library that estimates isotopic compositions through machine learning algorithms with user-defined variables. It includes dataset preprocessing, outlier detection, statistical analysis, feature selection, model validation and calibration and postprocessing. This tool has the flexibility to operate with discontinuous inputs in time and space. The automatic decision-making procedures are knitted in different stages of the algorithm, although it is possible to manually complete each step. The extensive output reports, figures and maps generated by Isocompy facilitate the comprehension of stable water isotope studies. WaterpyBal is another tool implemented in Python that generates spatial-temporal water balance models. The tool focuses on diffused precipitation and recharge modelling, considering the vertical water movement. The tool offers flexibility in terms of input data and modeling time interval, and integrates different stages of the water balance assessment such as spatial data interpolation, evaporation, evapotranspiration and infiltration calculation, taking into account the soil characteristics and urban water cycle parameters. WaterpyBal calculates the water budget parameters such as recharge, deficit and runoff in defined spatial-temporal spectrum and could be used in the post-processing stage to create maps, figures, datasheets and raster archives. This tool has a modular design with ensures the ability of further developments and add-ons. WaterpyBal Studio is the graphic user interface of WaterpyBal that facilitates the usage of this tool in water management projects.

In essence, these tools offer support for sustainable groundwater management projects that ensures reproducible research results in these environmental fields.

## Resumen

La modelación numérica de aguas subterráneas es esencial para una gestión sostenible del agua, ya que ayuda a comprender el comportamiento de los recursos de aguas subterráneas y su interacción con el agua superficial. Estos modelos pueden predecir los impactos de la manipulación del agua en el medio ambiente, ayudar en el desarrollo de escenarios de gestión sostenible y ayudar a comprender los impactos del cambio climático en los recursos de aguas subterráneas. La generación de modelos numéricos de aguas subterráneas requiere diversos datos de entrada y condiciones de contorno, que incluyen modelos geológicos, entrada de agua superficial y recarga de aguas subterráneas. Al combinar diferentes enfoques de modelado, se puede desarrollar una comprensión integral de los recursos de aguas subterráneas para apoyar la gestión sostenible de aguas subterráneas, la toma de decisiones y la planificación de recursos hídricos. Esta tesis se basa en herramientas de código abierto para ayudar en las diferentes etapas de los estudios de aguas subterráneas: modelado geológico 3D (Geopropy), modelado isotópico (Isocompy) y modelado espaciotemporal del balance hídrico del suelo (WaterpyBal y WaterpyBal Studio). Geopropy es un algoritmo de toma de decisiones implementado en Python, que genera secciones transversales 3D. Actúa como un agente inteligente que simula los pasos que realiza el geólogo en el proceso de creación de la sección transversal, combinado con decisiones basadas en datos. El algoritmo detecta zonas con más de una posible solución y, en función del nivel de complejidad (o preferencia del usuario), procede a las etapas automáticas, semiautomáticas o manuales. Isocompy es una biblioteca de Python que estima las composiciones isotópicas a través de algoritmos de aprendizaje automático con variables definidas por el usuario. Incluye pre procesamiento de conjuntos de datos, detección de valores atípicos, análisis estadístico, selección de características, validación y calibración del modelo y post procesamiento. Esta herramienta tiene la flexibilidad de operar con entradas discontinuas en el tiempo y el espacio. Los procedimientos automáticos de toma de decisiones están tejidos en diferentes etapas del algoritmo, aunque es posible completar manualmente cada paso. Los extensos informes, figuras y mapas generados por Isocompy facilitan la comprensión de los estudios de isótopos estables del agua. WaterpyBal es otra herramienta implementada en Python que genera modelos espaciotemporales de balance hídrico. La herramienta se centra en la modelación de la precipitación difusa y la recarga, considerando el movimiento vertical del agua. La herramienta ofrece flexibilidad en cuanto a los datos de entrada y el intervalo de tiempo de modelado e integra diferentes etapas de evaluación del balance hídrico, como la interpolación de datos espaciales, el cálculo de evaporación, evapotranspiración e infiltración, teniendo en cuenta las características del suelo y los parámetros del ciclo del agua urbana. WaterpyBal calcula los parámetros del presupuesto de agua, como la recarga, el déficit y la escorrentía, en un espectro espacial-temporal definido y se puede utilizar en la etapa de post-procesamiento para crear mapas, gráficos, hojas de datos y archivos raster. Esta herramienta

tiene un diseño modular que asegura la capacidad de futuros desarrollos y complementos. WaterpyBal Studio es la interfaz gráfica de usuario de WaterpyBal que facilita el uso de esta herramienta en proyectos de gestión del agua.

En esencia, estas herramientas ofrecen soporte para proyectos de gestión sostenible de aguas subterráneas que garantizan resultados de investigación reproducibles en estos campos ambientales.



## Resum

La modelització numèrica d'aigües subterrànies és essencial per a la gestió sostenible de l'aigua, ja que ajuda a comprendre el comportament dels recursos d'aigua subterrània i la seva interacció amb les aigües superficials. Aquests models poden predir els impactes de la manipulació de l'aigua en l'entorn, ajudar en el desenvolupament de escenaris de gestió sostenible i comprendre els efectes del canvi climàtic en els recursos d'aigua subterrània. La generació de models numèrics d'aigües subterrànies requereix diverses dades d'entrada i condicions límit, incloent models geològics, entrades d'aigua superficial i recàrrega d'aigua subterrània. A través de la combinació de diferents enfocaments de modelització, es pot desenvolupar una comprensió completa dels recursos d'aigua subterrània per donar suport a la gestió sostenible d'aigües subterrànies, la presa de decisions i la planificació dels recursos hídrics. Aquesta tesi es basa en eines de codi obert per ajudar en les diferents etapes d'estudis d'aigua subterrània: modelització geològica en 3D (Geopropy), modelització isotòpica (Isocompy) i modelització espacial-temporal del balanç hídric del sòl (WaterpyBal i WaterpyBal Studio). Geopropy és un algorisme de presa de decisions implementat en Python, que genera seccions transversals en 3D. Funciona com a agent intel·ligent que simula els passos que realitza el geòleg en el procés de creació de la secció transversal, enllaçat amb decisions basades en dades. L'algorisme detecta zones amb més d'una possible solució i, en funció del nivell de complexitat (o preferència de l'usuari), passa a les etapes automàtiques, semiautomàtiques o manuals. Isocompy és una biblioteca de Python que estima les composicions isotòpiques a través d'algoritmes d'aprenentatge automàtic amb variables definides per l'usuari. Inclou el pre-processament de conjunts de dades, la detecció de valors atípics, l'anàlisi estadística, la selecció de característiques, la validació i calibració del model i el post-processament. Aquesta eina té la flexibilitat per operar amb entrades discontinües en el temps i l'espai. Els procediments de presa de decisions automàtiques estan integrats en diferents etapes de l'algorisme, encara que és possible completar manualment cada pas. Els extensos informes de sortida, figures i mapes generats per Isocompy faciliten la comprensió dels estudis d'isòtops estables de l'aigua. WaterpyBal és una altra eina implementada en Python que genera models espai-temporals d'equilibri hídric. L'eina es centra en la modelització de la precipitació difusa i la recàrrega, tenint en compte el moviment vertical de l'aigua. L'eina ofereix flexibilitat en termes de dades d'entrada i interval de temps de modelització, i integra diferents etapes de l'avaluació de l'equilibri hídric, com ara la interpolació de dades espacials, el càlcul de l'evaporació, l'evapotranspiració i la infiltració, tenint en compte les característiques del sòl i els paràmetres del cicle urbà de l'aigua. WaterpyBal calcula els paràmetres del pressupost hídric, com ara la recàrrega, el dèficit i l'escorrentia en un espectre espai-temporal definit i es pot utilitzar en la fase de post-processament per crear mapes, gràfics, fulls de dades i arxius ràster. Aquesta eina té un disseny modular que garanteix la

capacitat de desenvolupaments i complements futurs. WaterpyBal Studio és la interfície gràfica d'usuari de WaterpyBal que facilita l'ús d'aquesta eina en projectes de gestió de l'aigua.

En essència, aquestes eines ofereixen suport per a projectes sostenibles de gestió d'aigües subterrànies que assegurin resultats de recerca reproduïbles en aquests camps ambientals.

# Contents

1	Introduction	1
1.2	MOTIVATION AND OBJECTIVES	2
1.3	THESIS OUTLINE	5
2	An automatic geological 3D cross-section generator: Geopropy, an open-source library	7
2.1	INTRODUCTION	8
2.2	METHODS	10
2.2.1	General definitions	10
2.2.2	Database	11
2.2.3	Steps of drawing a cross-section	12
2.3	THE GEOPROPY CODE	13
2.3.1	Geopropy architecture	14
2.3.2	Geological input data and HYDOR database modifications	14
2.3.3	Geopropy workflow	17
2.3.4	Output and visualization	21
2.4	APPLICATION TO SYNTHETIC DATASETS	21
2.4.1	Synthetic database 1	22
2.4.2	Synthetic database 2	23
2.4.3	Synthetic database 3	24
2.5	DISCUSSION AND CONCLUSION	28
2.6	SOFTWARE AND DATA AVAILABILITY	29
2.6.1	GEOPROPY library information:	29
2.6.2	Synthetic examples	30
3	An open source Python library for environmental isotopic modelling	31
3.1	INTRODUCTION	32
3.2	METHODS	33
3.3	UNDER THE HOOD OF ISOCOMPY	37

3.3.1	Isocompy workflow	37
3.3.2	Isocompy architecture	39
3.3.3	Outputs	43
3.4	APPLICATION TO THE EXAMPLE OF SALAR DE ATACAMA	44
3.4.1	Input data	46
3.4.2	Implementation	46
3.4.3	Results and discussion	49
3.5	CONCLUSION	55
3.6	SOFTWARE AND DATA AVAILABILITY	56
3.6.1	Isocompy library information	56
3.6.2	Application on Salar de Atacama	56
4	An open source Python library for water balance modelling	57
4.1	INTRODUCTION	58
4.2	METHODS	59
4.2.1	Soil Water Balance	60
4.2.2	Infiltration and Runoff	60
4.2.3	Soil Water Reserve	63
4.2.4	Potential Evapotranspiration	63
4.2.5	Recharge, Real Evapotranspiration, and Deficit	64
4.2.6	Urban Cycle	65
4.3	UNDER THE HOOD:	66
4.3.1	dataset_gen	67
4.3.2	variable_management	68
4.3.3	SWR	68
4.3.4	PET	68
4.3.5	Infiltration, Urban_composite_CN, and Urban_cycle	68
4.3.6	Balance	69

4.3.7	post_process	69
4.4	THE WORKFLOW	69
4.5	GRAPHIC USER INTERFACE OF THE WATERPYBAL LIBRARY	71
4.6	SYNTHETIC EXAMPLE OF THE WATERPYBAL APPLICATION	71
4.6.1	Designed Study Areas and Available Data	71
4.6.2	SWB Input Data Preparation:	74
4.6.3	WaterpyBal Application:	74
4.6.4	Example Results and Discussion:	75
4.7	CONCLUSIONS	77
4.8	SOFTWARE AND DATA AVAILABILITY	78
5	Conclusions	79
	References	82
	Appendices	103
A.	Additional Figures of Chapter 2	103
B.	Synthetic Datasets Tables of Chapter 2	106
C.	Listings of Chapter 2	116
D.	Urban cycle calculations – Chapter 4	118
E.	Empirical methods equations and parameters – Chapter 4	120
F.	Additional figures – Chapter 4	121
G.	List of Scientific and Technical Production	123
H.	Cover of the scientific articles	128

## List of figures

Figure 2.1 Geological units from oldest to youngest: c, b, a. Higher and lower priority numbers indicate younger and older unit contacts, respectively...	13
Figure 2.2 Geological section of the HYDOR database. The orange tables are the original ones in the database. For more information...	15
Figure 2.3 Simplified schematic Geopropy decision-making algorithm. CZ: Critical Zone...	18
Figure 2.4 Geopropy algorithm steps used to generate a synthetic cross-section (CS). Red dots: fault points. Orange dots: conformable contacts. Two coloured dots: unconformable contacts...	20
Figure 2.5 First synthetic example. The colours and letters correspond to geological units. The example contains a normal fault...	23
Figure 2.6 Second synthetic example. The colours and letters correspond to geological units. This includes a normal fault and an intrusion...	24
Figure 2.7 Raw data of the third synthetic example. The circles show the available surface data. The coloured circles show information about the geological units...	25
Figure 2.8 1 and 2: Two possible interpretations of zone <i>iii</i> drawn by the geologist. 3 and 4: Two possible interpretations of zone <i>iv</i> drawn by the geologist...	26
Figure 2.9 Results of the third synthetic example. The colours and letters correspond to geological units. 1: Raw data illustrated in the coloured columns and the cross-section completed by the geologist...	27
Figure 3.1 Workflow scheme for estimating isotopic values by using two independent parameters that are available in different locations than the isotopic measurements...	35
Figure 3.2 Scheme of the Isocompy workflow utilized to design the Isocompy architecture. It consists of data preparation (red boxes) and two main stages...	38
Figure 3.3 Yellow, green and violet boxes show the techniques used in the statistical analysis step, the regression methods available in Isocompy...	38
Figure 3.4 The Isocompy algorithm architecture. It contains 6 classes and 18 methods.	40
Figure 3.5 Feature selection flowchart of the model class. Red lines indicate false arguments.	41
Figure 3.6 Workflow of the model regression, model validation, model calibration and best model selection processes. Black dots show that these processes are performed for each regression method selected.	41
Figure 3.7 Workflow of the evaluation class for estimating the second-stage regressions.	42
Figure 3.8 Screenshots of the outputs generated by Isocompy. Top left: An example report. Bottom left: Partial dependency plots of the selected features...	44
Figure 3.9 Left: location map of the study area in South America with published isotopic precipitation data (red circles) and automatic weather stations that monitor temperature (crosses)...	45

Figure 3.10 Isocompy data preparation. Location information (X, Y: coordinates; Z: altitude) is used to calculate the feature information in these positions...	47
Figure 3.11 Stage-one estimation models, estimator and partial dependency plots.	48
Figure 3.12 Stage-one estimation calculations (line 2). Stage-two model argument definitions (lines 7-10). Stage-two model execution (line 16). Statistical reports and plots (lines 26-27).	49
Figure 3.13 Plots of the estimated-versus-observed values generated by Isocompy for temperature, precipitation and relative humidity in January, February and March.	51
Figure 3.14 Maps of the temperature, precipitation, relative humidity and $\delta^{18}\text{O}$ values of precipitation estimated by Isocompy in January, February and March in the Salar de Atacama basin.	52
Figure 3.15 Top left and top right: estimations versus the measurements of $\delta^{18}\text{O}$ and $\delta^2\text{H}$ , respectively. Bottom: plots of the estimated (circles) and observed (triangles) $\delta^{18}\text{O}$ versus $\delta^2\text{H}$ values...	54
Figure 4.1 The general scheme to calculate the SWB. PET: Potential Evapotranspiration, RET: Real Evapotranspiration...	60
Figure 4.2 Simplified scheme of the urban water cycle. The abbreviations are explained in this section.	66
Figure 4.3 Simplified scheme of the main classes and methods of WaterpyBal. The colors used for each class correspond to the process with the same color in Fig. 1.	67
Figure 4.4 Scheme of the input types at each step of SWB calculation and modular structure of WaterpyBal. The colors used in each process correspond to the same stage...	67
Figure 4.5 Synthetic example of a study area. The study area is divided into six regions with diverse characteristics. Regions 2 and 3 are urban areas...	72
Figure 4.6 (a) Monthly average precipitation of nine meteorological stations. (b) Average temperature measurements for nine meteorological stations.	73
Figure 4.7 Precipitation, recharge, and RET values in regions 1 and 4. Values in mm/day.	75
Figure 4.8 Dashed lines show the annual recharge calculated by WaterpyBal, and the vertical bars show the two standard deviations of the average of the empirical methods...	76
Figure 4.9 (a, b) Precipitation, recharge, and evaporation in regions 2 and 3, respectively. (c) Recharge evolution by changing the UNUA value in an urban area. Values in mm/day.	77

## List of tables

Table 2.1 Example of a chronological data table. The unit contact types can be <i>conformity</i> , <i>unconformity</i> , <i>intrusion</i> or <i>fault</i> . _____	16
Table 2.2 Example of a Fault_data table. The <i>Borehole_ID</i> field must be selected from a drop-down list derived from the borehole data table. _____	16
Table 2.3 Example of a ground surface data table. The grey row indicates a ground surface point with only geospatial information available. The first two rows show outcroppings with defined unit contacts. ____	16
Table 3.1 Results of the first-stage statistical analysis and models per month. The bold p values denote significant parameters (<0.05). _____	50
Table 3.2 The VIF values and correlation coefficients of the second-stage input features. VIF_init and VIF_fin show the initial and final VIF values, respectively... _____	53
Table 4.1 Soil characteristics in six regions of the synthetic example. Zones 2 and 3, marked by *, are urban zones... _____	73
Table 4.2 Urban cycle parameters in regions 2 and 3. _____	74
Table 4.3 Annual recharge calculated by WaterpyBal and the statistics of the annual recharge calculated by four empirical methods. Values in mm/year... _____	76



# **1 Introduction**

## **1.2 MOTIVATION AND OBJECTIVES**

Water is a vital resource for life on Earth, and its availability is becoming more limited due to droughts and pollution. As the world's population continues to grow, the demand for water is increasing, and effective water management becomes increasingly important <sup>1</sup>. Poor water management practices exacerbate water scarcity, which is a significant problem in many parts of the world. Climate change is affecting the availability of water, with changing weather patterns and increasing temperatures leading to more frequent and severe droughts <sup>2</sup>.

Groundwater resources play a crucial role in water management, particularly in areas where surface water is limited or unreliable. Groundwater is an important source of water and it can provide a reliable supply of water even during periods of drought or low rainfall. However, the management of groundwater resources can be challenging, as overuse or contamination of the resource can have negative environmental impacts. Effective groundwater management requires a holistic approach that balances the need for water with the need to protect the resource for future generations <sup>3</sup>.

Modern groundwater management has evolved to become more quantitative, relying on data-driven assessments and advanced modeling techniques. This approach allows for a more accurate assessment of the resource and can help to inform management strategies that are based on sound science and accurate information <sup>4</sup>. Groundwater management also involves a greater focus on sustainability and the protection of the resource for future generations, which includes the development of strategies to minimize the environmental impacts of groundwater use <sup>5</sup>.

Conceptual groundwater models and groundwater numerical modeling are essential tools for effective groundwater resource management. Conceptual models provide simplified representations of the groundwater system to identify potential sources of contamination and predict impacts of groundwater withdrawals <sup>6</sup>. Groundwater numerical modeling involves the use of mathematical equations to simulate the behavior of the groundwater system and is particularly useful for simulating complex hydrogeological conditions. The benefits of using these models include a more accurate and comprehensive understanding of the groundwater system, identifying potential risks, evaluating different management strategies, and supporting stakeholder engagement and decision-making. However, challenges associated with these models include the need for accurate data and potential uncertainty and error <sup>7</sup>.

Groundwater numerical modeling is crucial for sustainable water management as it helps understand the behavior of groundwater resources and predicts the impacts of water manipulation on the environment and human activities such as mining and construction. Additionally, it aids in understanding the impacts of climate change on groundwater resources, identifying and mitigating these impacts, and predicting future

water availability by assessing groundwater recharge and discharge rates. Generating numerical groundwater models need various input data and boundary conditions related with geological models, water origin and quantitative groundwater recharge being some of the vital aspects <sup>8</sup>.

Geological models can provide the hydrogeological framework for numerical groundwater models by defining the geometry, boundary conditions and hydraulic properties of the subsurface aquifer system. They can also help to identify the groundwater flow paths and direction of water movement through the aquifer, which are important inputs for numerical groundwater models. These models are an essential part of the earth sciences, enabling geologists to understand and interpret the physical characteristics and properties of the subsurface. The creation of a geological model involves analyzing various data sets and combining them to create a three-dimensional representation of the subsurface. Such models are used in various applications, including resource exploration and development, environmental management, and risk assessment <sup>9</sup>. In general, geological models can be described from implicit or explicit points of view. Implicit modeling includes data-driven methods that use datasets derived from measured features and algorithms. The main advantage of implicit modeling is its ability to generate models quickly and efficiently, even in areas where data is scarce or incomplete. On the other hand, explicit modeling mostly relies on expert opinion, experience, and interpretation, where the expert usually builds cross-sections, surfaces or volumes by interpolating the accessible data. Explicit modeling allows for more accurate and detailed representations of the subsurface but can be time-consuming and expensive, especially in areas where data is limited <sup>10</sup>.

Despite the advancements in implicit geological modeling tools, approaching 3D modeling and cross-section generation by relying on geologists' experience is still an acceptable outcome in many scenarios, which is why it is still used in numerous ongoing projects in industry and the academic community alike. However, there is a need for more efficient and effective methods that can combine the advantages of both implicit and explicit modeling approaches <sup>11-13</sup>.

Water isotope models can help to identify the sources of the surface water and groundwater recharge, which can provide critical input data for numerical groundwater models. The stable water isotopes,  $^{18}\text{O}$  and  $^2\text{H}$ , provide information about different aspects of the hydrological cycle <sup>14</sup>. They are commonly used in meteorological and hydrological studies to identify precipitation origin, study local effects in soil water balance studies, determine the relative contributions of water with different sources in a water body, describe the process of aquifer recharge and characterization, and investigate various aspects of runoff and stream flow generation <sup>15,16</sup>. All of these features are essential for optimal and sustainable water resource management.

There are two main approaches for studying the global distribution of isotopic composition in precipitation: isotope-enabled atmospheric general circulation models (IGCMs) and regression statistics-based approaches <sup>17</sup>. IGCMs are numerical models that consider various physical processes, including water isotopes, while regression statistics-based models are useful for identifying possible processes based on isotopic signatures <sup>18</sup>.

The statistical models have limitations, such as the lack of a standalone tool for determining input features and databases for developing a statistical isotopic model, scarce isotopic data, different types of isotopic samples, and neglecting underlying processes of water isotopic signature due to the use of simple linear models <sup>17,19,20</sup>.

Water balance models can simulate the interactions between surface water and groundwater, which are important for numerical groundwater models that need to account for recharge from diverse water sources. Groundwater recharge refers to the process by which precipitation and other water resources enter groundwater reservoirs, or aquifers <sup>21</sup>. Recharge assessment also allows experts to determine the soil water balance in a specific area and take appropriate measures to ensure sustainable groundwater management. Water balance models can also simulate the rate of deficit, runoff and evapotranspiration, which is a vital input for estimating the water balance of an area and specifically the recharge into the aquifer <sup>22</sup>. There are various methods and computer programs available to estimate soil water balance parameters, and the selection of the appropriate tool depends on the objectives of the study and data availability <sup>23</sup>. However, many existing computer programs for recharge assessment have some limitations that have not been fully addressed. These limitations include not accepting spatial data, the use of a strict time window, the inflexibility of input data, not taking into account the complexity of urban infrastructure, limited options for water recharge calculation methods, not being open-source, the need for knowledge of scripting languages, limited outputs, not incorporating widely-used databases and limitations in incorporating recent advancements in computation power and machine learning algorithms <sup>24-33</sup>.

Overall, by combining these different modeling approaches, a more comprehensive and quantitative understanding of the of the groundwater resources can be developed which can be used to generate more accurate and reliable conceptual and numerical groundwater models. This, in turn, can be used to support sustainable groundwater management, evaluate the impacts of potential groundwater use scenarios, and support water resource planning and decision-making.

The general objective of this thesis is to acquire reliable methods of water resource management by automatizing geological and hydrogeological calculations of processes. This is achieved by creating open-source tools that would aid the user in the different stages of groundwater studies, having in mind the

structural algorithm design to ensure the future developments and to contribute to the reproducible research. This objective will ensure the numerical predictivity of groundwater management. The specific goals can be detailed as follows:

- Create a hybrid data-knowledge driven geological 3D modelling tool that aims to automate the process of creating geological cross-sections based on the complexity of the input data that aids the user in implicit geological modeling. The goal on this tool is to follow the same thinking steps as the geologist rather than focusing on the mathematical interpretation of the input data.
- Build a tool for statistical-regression based isotopic modelling of water stable isotopes using extensive statistical analysis, decision-making procedures and modern estimation algorithms to maximize the automation of the process of creating the isotopic models. The goal is to offer the capability of operating with discontinuous inputs in time and space and the quantity of the input features. This tool has to create logs and reports that allows the user to investigate the underlying processes from the preprocessing to postprocessing stage while maintaining the ease of use for the user with diverse levels of experience.
- Construct a tool for water balance modelling, considering spatial and temporal variations of data, focusing on diffused precipitation and recharge. The goal of this tool is to offer flexibility in terms of input data and modeling time interval while integrating different stages of the water balance assessment such as spatial data interpolation, evaporation, evapotranspiration and infiltration calculation, taking into account the soil characteristics and urban water cycle parameters and post-processing. This tool has to be complemented by a graphic user interface that facilitates the usage of this tool in water management projects for the user with limited coding experience.

### **1.3 THESIS OUTLINE**

This thesis includes five chapters: introduction (this chapter), three main chapters which correspond to the three specific objectives and numerical tools, and the main conclusions. The thesis is based on the scientific articles that are published or submitted to international journals. The references to the publications are included in the beginning of each chapter. Each of these chapters/publications contain their own introduction to the problem as well as the methodology employed.

*Chapter two* is dedicated to Geopropy, a tool that facilitates geological modeling. It performs as an intelligent agent that simulates the steps taken by the geologist in the process of creating cross-sections,

coupled with data-driven decisions. It is an open-source Python-based library that generates consistent 3D geological cross-sections by hybrid implicit-explicit methods. The aim of Geopropy is to automate parts of the geologist's workflow to support the decision-making process, accompanied by data-driven decisions. In conclusion, Geopropy enables more efficient and effective decision-making, allowing for the creation of more accurate and detailed geological models.

*The main focus of the third chapter* is on an open source, Python-based, multistage isotopic composition analysis and modelling library named as Isocompy. Its main objectives are to integrate the diverse steps of stable statistical isotope modelling, incorporating novel data management and machine learning regression methods, and exhibiting flexibility regarding to the available input data. In general, it is thought to be intuitive and user-friendly, and generates reports and figures in every step. Functionality of Isocompy is demonstrated using an application example involving the meteorological features and isotopic composition of precipitation in N Chile.

*Chapter four focuses on* a tool for water balance modeling which is a key in groundwater management projects. WaterpyBal is an open-source modular python library that supports different stages of soil water balance assessment. The library incorporates the principles of hydrological/watershed modeling methods and offers flexibility in input data, time interval, and spatial-temporal properties. Moreover, it provides a collection of tools to facilitate different stages of the study. The library also includes a well-known database as the core dataset, with the ability to integrate a broad range of information that is supported by a wide series of programs. WaterpyBal can account for the complexities of urban infrastructure, which is essential for regional studies. The library also offers a wide range of water recharge calculation methods. The library's modular design allows for future developments and contributions to reproducible research, making it an essential tool for groundwater management and water management projects.

WaterpyBal is accompanied by WaterpyBal Studio, a graphic user interface that allows users to use the library without the need for scripting knowledge. The functionality of the library is demonstrated with a synthetic example, highlighting the flexibility and effectiveness of the library.

*Chapter five* summarizes the main conclusions of the thesis and the recommendations for future works

In addition, there are eight appendices that summarize the following information:

Appendices A to F: additional information on the second and fourth chapter.

Appendix G: List of scientific and technical production.

Appendix H: Cover of the scientific articles.

## **2 An automatic geological 3D cross-section generator: Geopropy, an open-source library**

This chapter is based on:

**Hassanzadeh, A.,** Vázquez-Suñé, E., Corbella, M. & Criollo, R. An automatic geological 3D cross-section generator: Geopropy, an open-source library. *Environmental Modelling & Software* 149, 105309 (2022). DOI: [10.1016/j.envsoft.2022.105309](https://doi.org/10.1016/j.envsoft.2022.105309).

## 2.1 INTRODUCTION

Obtaining 3D subsurface geological models is of great importance in a wide variety of geoscience studies. These models represent superficial and underground geological structures and the distribution of geological units, and their purpose is to illustrate the existing underground conditions. Subsurface geological understanding and models are essential aspects of earth-related industrial projects and academic investigations, from mining and petroleum to hydrogeology and environmental studies <sup>34–37</sup>.

In general, geological models can be described from implicit or explicit points of view. Implicit modelling includes data-driven methods that use datasets derived from measured features and algorithms. Explicit modelling mostly relies on expert opinion, experience and interpretation, where the expert usually builds cross-sections, surfaces or volumes by interpolating the accessible data <sup>10</sup>.

To date, various works have presented and discussed the use of implicit methods, including cokriging-based modelling <sup>38–41</sup>, volume-based modelling <sup>42,43</sup>, kinematic modelling <sup>44,45</sup> and others <sup>46,47</sup>. In addition, open-source software such as Noddy/pynoddy <sup>48,49</sup> and GemPy <sup>50,51</sup> and commercial software such as GDM Suite <sup>52</sup>, MOVE <sup>53</sup>, Vulcan 3D/EUREKA <sup>54</sup>, Leapfrog <sup>55</sup> and Oasis Montaj <sup>56</sup> are powered by these methods. Most of these computer programs contain more than one method, and although they are known as implicit modelers, they benefit from explicit modelling in some stages of geological model evaluation.

For explicit modelling, there are computer programs that support experts in generating cross-sections. Some examples of open-source software are HEROS <sup>57</sup> and HEROS 3D <sup>58</sup>, the Midvatten QGIS plugin <sup>59</sup>, Grass GIS <sup>60</sup>, and Geomodelr <sup>61</sup>. There are also commercial software programs that support explicit modelling, such as RockWare GIS <sup>62</sup>, Surpac <sup>63</sup>, Datamine <sup>64</sup>, GeoScene3D <sup>65</sup>, BGS Groundhog Desktop <sup>66</sup>, and GSI3D <sup>67</sup>. One can also use other characteristics to categorize these programs: the amount or type of information used to build a model, the proportion of knowledge or data that drives the model, the level of automation, the focused scope, the computational performance or the degree of specialty needed to use them <sup>68</sup>.

Despite the advancements in implicit geological modelling tools, approaching 3D modelling and cross-section generation by relying on geologists' experience is an acceptable outcome in many scenarios, which is why it is still used in numerous ongoing projects in industry and the academic community alike. Despite the diversity of methods and computer programs available, we believe that there are some shortcomings that have not been fully addressed to date:

- (1) One common key aspect of generating geological models is database integration. Modifying databases from well-known formats to a specifically structured shape could be time consuming, specifically in complex projects with a considerable amount of data <sup>57</sup>.



- (2) Projects normally need various cross-sections, so generating them could be a time-consuming procedure.
- (3) In explicit modelling, the lack of geologist experience could result in outcomes that are far from the reality of the region <sup>11</sup>.
- (4) There are many scenarios with more than one possible outcome. Freedom in the style and bias of each expert could play a role in determining the final cross-sections, which will result in nonunique solutions <sup>12</sup>.

Points three and four may result in inconsistent models, especially in projects where more than one geologist is working on cross-sections <sup>13</sup>, which could potentially result in decreasing the feasibility of using the explicit models.

- (5) The lack of experience of the geologist with mathematical modelling methods, specifically in moderately complex scenarios, could make it challenging to follow the steps and interpret the cross-sections.
- (6) Some advanced packages do not provide extensive information about the procedures and methods undertaken without purchasing a licence (e.g., Petrel <sup>69</sup>, GeoModeller <sup>70</sup>, Go-CAD <sup>71</sup>).
- (7) The rapid development of artificial intelligence and machine learning libraries (e.g., TensorFlow (Abadi et al.), Scikit-learn <sup>73</sup>, Pytorch <sup>74</sup>, Pymc <sup>75</sup>) suggests a potential for expandability in geological modelling software. However, some existing computer programs are limited in this matter (e.g., EarthVision <sup>76</sup>, RockWorks, Strater <sup>77</sup>), and the developments are mostly focused on reservoir properties <sup>78-80</sup> or environmental modelling <sup>3,81-83</sup>.

Developing open-source software programs based on widely used programming languages brings the possibility of i) understanding and replicating the undertaken procedures, ii) customizing the code for further analysis or different purposes in case of need and *iii*) dynamically improving the software by implementing the latest version of machine learning and decision-making methods that have already been developed by third parties <sup>84</sup>. Moreover, using a standardized, known database could ease the process of introducing data to the software <sup>85</sup>. More importantly, to the best of our knowledge, there is no algorithm that aims to reproduce the steps of creating a cross-section by the same approach used by a geologist, which we think is worth considering, trying and developing for research and industry.

In this work, we address some of the shortcomings described above by developing Geopropy, an open-source Python-based <sup>86</sup> library, to generate consistent 3D geological cross-sections by hybrid implicit-explicit methods. Geopropy completely or partially automates the workflow of the geologist to support the

decision-making process, accompanied by data-driven decisions. The objectives of Geopropy are as follows:

1. It aims to identify the uncertainty caused by complex structures or a lack of data by detecting zones with more than one possible outcome and facilitating the decision-making process in these zones for the user.
2. The algorithm tries to act as an intelligent agent that emulates the decision-making steps of a geologist by following the geological complexity with three different degrees of freedom:
  - a) If the algorithm detects that the available information results in a unique outcome, the cross-sections will be generated automatically.
  - b) If the algorithm determines that there is more than one possible outcome for the cross-section, it proceeds to the semiautomatic stage, in which it asks for decisions on how to complete the geological unit contacts.
  - c) If new geospatial information or more complex decisions are needed to complete the cross-section, the algorithm enters the manual stage to complete the cross-section.

Although fundamental uncertainties in geological and environmental models and various studies around them are of great importance<sup>87-92</sup>, uncertainty is currently not reflected in Geopropy.

In the next sections, we will describe the methods applied in this work, followed by the details of the Geopropy code and the database modifications. To evaluate the algorithm, we will demonstrate the application of the algorithm on three synthetic datasets and compare it with explicit methods.

## **2.2 METHODS**

In this section, we will clarify geological unit contact types to unify the definitions for the machine and the geologist, and then we will explain the database that is used in Geopropy. We will discuss the extraction of the logical steps for cross-section generation based on the geologist's decision-making steps.

### **2.2.1 General definitions**

To translate the thinking of the geologist to a machine language, unified assumptions are essential. With that in mind, various geological structures are defined so that they are distinguishable by the geologist and Geopropy simultaneously:

- i. A conformity (conformable contact) represents continual, uninterrupted deposition and accumulation of sedimentary rocks without a gap in the geologic record.
- ii. An unconformity (unconformable contact) is a type of stratigraphic contact between a specific sedimentary material in the top layer and other older units that are not chronologically correlative as the bottom layer. This contact may cut previous surfaces. Unconformity contacts include disconformity, nonconformity, angular unconformity and paraconformity contact types. Geopropy recognizes all unconformity contact types as the same group.
- iii. An intrusion is an intrusive body of rock that solidifies inside existing geological units. An intrusion has priority with respect to the units it intrudes into, since it forms after them.
- iv. Faults are defined as surfaces bounding any kind of geological unit that cuts any other previous unit, and they imply a relative movement between the two blocks at each side of the surface <sup>93</sup>.

Using these definitions, Geopropy determines the zones that contain complex structures such as reverse faults, folds and repetition of the same geological unit in a borehole. We name these regions “critical zones” and the respective points “critical points”. These critical zones could be the source of multiple cross-sectional outcomes for a specific dataset.

### **2.2.2 Database**

An increasing volume of data in earth science demands a database with the ability to integrate various types of information for further practice <sup>94</sup>. To ensure a dynamic workflow and the replicability of the analysis, it is advised to manage the data collected in a standard database. In addition, flexibility in use with several tools and user-friendliness are other properties of an adequate database. We use the HYDOR data platform as the main Geopropy database to continue the developments made by Velasco (2013, Alcaraz (2016) and Criollo (2019). HYDOR was built as a personal geodatabase (ESRI) that can be used as a relational Microsoft Database, which facilitates data management with a user-friendly interface such as Microsoft Access software <sup>98</sup>, the ArcGIS platform <sup>99</sup> or QGIS <sup>100</sup>. HYDOR has been widely used and tested <sup>94</sup>, its data tables are homogenized and it cross analyses different types of information such as geology, hydrogeology, and environmental information, which could improve the validation of conceptual models. In addition, there are tools such as HEROS, HEROS 3D, QUIMET <sup>101</sup>, ArcArAz <sup>96</sup>, Metrogeother <sup>102</sup>, MJ-Pumpit and HYYH <sup>103</sup> that have already been designed based on this database. The HYDOR design follows established guidelines such as OneGeology, 2013; OGC Standards, 2011; OGC WaterML 2.0, 2012; and INSPIRE, 2013.

### 2.2.3 Steps of drawing a cross-section

To map out the automation of cross-section construction by following the same strategy a geologist would, we performed an exercise in which several geologists of diverse levels of expertise (4 university professors in the geological modelling field, 5 bachelor's students of geology and 2 PhD candidates in the geoscience field) drew a cross-section based on the same dataset (refer to section 2.4.1) so that they could describe their thinking steps and considerations while performing a geological interpretation. In general, they proceeded with the following steps: 1) Identification of the geological units present in the area and their relative ages and interpretation of the types of contacts that separate the units in the boreholes. 2) Consideration of the outcroppings and ground surface data and flagging the critical points that can affect the layer contacts. 3) Focusing on complex structured zones such as folds and reverse faults, keeping in mind the geological maps, relative age of the layers and critical ground surface points. At this stage, the geologist may have one or more possible outcomes to complete the contacts, respecting all boundary conditions. The contacts will be determined by considering all the information, the experience of the geologist and personal preference. Then, the geologist completes the cross-section by depicting relatively simple unit contacts.

There are some simple and logical assumptions extracted from the steps that a geologist considers in constructing the cross-section:

- If one contact of two specific geological units (for example, the a/b contact in boreholes 1 and 2) exists in two consecutive boreholes, the boundary between the two geological units (orange and black lines) must be defined by either a straight or a curved line.
- In the case of the intersection of two geological contacts, following Hutton's cross-cutting relation principle<sup>108</sup>, the younger contact has priority over the older contact since it is derived from a newer geological event (Fig. 1, blue dot).
- If there is no unit contact information that forces the position of the contact, the geologist obtains the apparent dip angle and tries to complete the contact using it or tries to continue the unit contact with the same angle (if this contact unit already exists in the previous boreholes). It is possible that using these two options will result in an inconsistent unit contact that cannot be verified by other available information or angular data that are not available. The dashed red line and dashed brown line in Fig. 1 are examples of drawing a line by using the apparent dip angle and by using the same angle as the yellow solid line, respectively, and none of them can be validated by existing geological units. In this case, the personal preference of the geologist plays an important role in determining the unit contact. The geologist could connect it to the younger contact at any point in the form of a curved or a straight line. The blue line in Fig. 1 shows an example of a contact that could be validated by existing information.

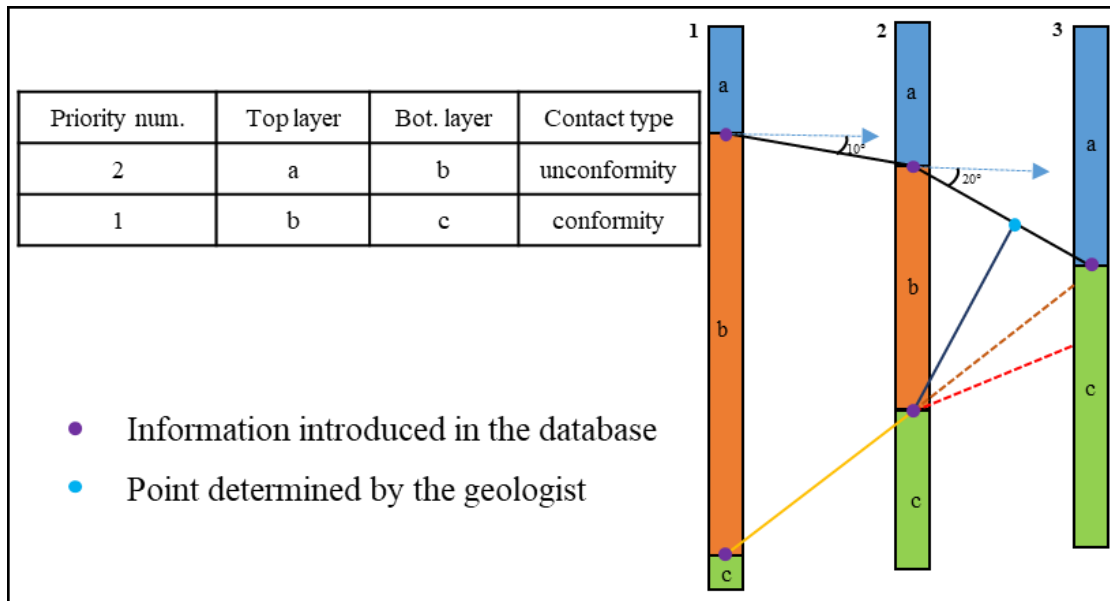


Figure 1 Geological units from oldest to youngest: c, b, a. Higher and lower priority numbers indicate younger and older unit contacts, respectively. The angles of the lines between the violet points are derived from the available geological unit data. The orange line shows a simple conformable contact, and the black lines show an unconformable contact. The dashed lines indicate inconsistent results. The dashed red line is derived from the apparent dip angle, and the dashed brown line is the result of continuing the orange line at the same angle. The blue dot illustrates a point that is chosen by the geologist based on personal preferences, which results in a consistent contact unit.

In the next sections, we will discuss the different aspects of the Geopropy code structure, followed by an application of the algorithm on three synthetic datasets to evaluate the functionality of the code by comparing it to the explicit cross-sections drawn by the geologist.

### 2.3 THE GEOPROPY CODE

The program is a hybrid implicit-explicit type, as it is designed to proceed automatically whenever possible, that is, whenever there is only one possible solution with the data available. However, it switches to a manual mode or asks for help from a geologist when there is more than one outcome. In this section, we will describe the architecture of Geopropy, followed by explaining the modifications of the HYDOR database and required input data. We will discuss the workflow of the algorithm in cross-sections containing simple and complex geological structures with an example, and we will show the available outputs and visualization options.

### **2.3.1 Geopropy architecture**

As an intelligent agent that tries to simulate the same stages as a skilled geologist, Geopropy is programmed to proceed as follows:

First, it extracts the geological units' contact points. Then, it examines the ground surface points and outcroppings to generate the ground surface contacts and to determine critical ground surface points. Afterwards, it analyses the contact points to detect repetitions of units in the same borehole, polarities, faults, the possibility of existing folds and the locations of faults and marks the critical zones. Then, for each critical zone, the algorithm proceeds to the automatic stage if there is just one possible solution. If there is more than one potential outcome, the algorithm asks the user in two substages to identify the desirable scenario in the critical zone, primarily by asking the user to choose the preferable solution between the existing options. If this information does not result in a unique outcome, the algorithm proceeds to the manual stage, where the user has to introduce more information. The workflow of the critical zones will be explained in the next sections in detail.

After analysing and determining the state of the contacts in the critical zones and accounting for the temporal sequence of geological events, the rest of the unit contacts will be completed based on the mentioned guidelines. It should be noted that it is possible to modify the cross-section in a postprocessing stage with any software that is compatible with 3D shape files.

### **2.3.2 Geological input data and HYDOR database modifications**

The geological dataset in HYDOR includes information related to borehole properties, samples, lithology and definitions of geological units and subunits. Geopropy uses lithological and borehole data. The lithology data table contains the borehole sample depth and geological units, whereas the borehole data table contains information about the locations of the boreholes (Fig. 2).

One of the requirements of processing the geological data similarly to using the criteria of a geologist is to have the same level of information available to the user in the decision-making procedure. Therefore, chronological data, fault data and ground surface data tables are added to the geological section of HYDOR to fulfill this need (Fig. 2, green tables).



Table 1 Example of a chronological data table. The unit contact types can be *conformity*, *unconformity*, *intrusion* or *fault*.

<i>priority_number</i>	<i>bottom_layer</i>	<i>top_layer</i>	<i>type</i>	<i>preferred_angle</i>
1	a	b	Conformity	
2		d	Intrusion	
3			Fault	
4		n	Unconformity	45

The fault data table (Table 2) must include the *priority\_number* that is already assigned in the chronological data table and the borehole ID and the elevation of the fault inside the borehole in the *Borehole\_ID* and *Elevation* fields, respectively. It is also optional to add an apparent fault dip in case of availability.

Table 2 Example of a Fault\_data table. The *Borehole\_ID* field must be selected from a drop-down list derived from the borehole data table.

<i>priority_number</i>	<i>borehole_id</i>	<i>elevation</i>	<i>preferred_angle</i>	<i>type</i>
5	119	20	45	fault

The ground surface data table (Table 3) consists of two groups of points: the points with only geospatial information (Table 3, grey row) and the outcroppings or unit contacts that are extracted from geological maps or field surveys. In the latter, the priority number of the unit contact point must be identified.

Fields *X*, *Y* and *Z* define the geospatial information of the sample. *Polarity* is an optional field that can be *Normal* or *Reverse*. *Reverse* polarity occurs when the unit contact defined in *priority\_number* is observed but the sequence of the unit is reversed. *angle* is an optional field to add the apparent dip of the unit contact.

Table 3 Example of a ground surface data table. The grey row indicates a ground surface point with only geospatial information available. The first two rows show outcroppings with defined unit contacts.

<i>x</i>	<i>y</i>	<i>z</i>	<i>priority_number</i>	<i>type</i>	<i>polarity</i>	<i>angle</i>
159	159	0	2	Topography	normal	
165	165	0	3	Topography	normal	0
260	260	0		Topography		



### 2.3.3 Geopropy workflow

The programming language used for the code presented is Python. Geopropy can be executed on Windows and Linux OS-based computers. The code will interact with databases in personal spatial geodatabase (ESRI) format.

The first step of running Geopropy is to complete the specific HYDOR database, as mentioned in section 2.3.2. Once the main function is executed, Geopropy reads the input information, extracting the outcroppings, ground surface, lithological unit contacts and chronological information. Then, the tool identifies 2D planes between every two consecutive boreholes and defines the boundary conditions respecting the arguments chosen by the user. Fig. 3 shows the simplified schematic decision-making algorithm of Geopropy.

Next, accounting for the relative age of the unit contacts, Geopropy identifies the zones defined as critical (with more than one possible scenario) in the ground surface and subsurface to complete the geological unit contact lines. The state of the critical zones determines the complexity of the process. The yellow box in Fig. 3 shows the procedure for each critical zone. If there is no identified critical zone, Geopropy passes directly to the automatic stage.

If there are critical zones, Geopropy first processes each zone in the semiautomatic stage. Some information will be shown to the user: contact unit info, critical zone boundaries, faults that are related to the unit contacts or exist within the critical zone, the point table that contains the specific point IDs for each critical point and the instruction to choose between points or to proceed to the manual stage. In the semiautomatic stage, the user can choose the sequence of points in the critical zone that have to be connected in the form of a Python list. If the contact in the critical zone has more than one connected piece, each piece can be separated by determining the 'SEPARATE' keyword in the list. (See example C.1 in the Appendices.) The semiautomatic stage is accompanied by a visual representation of the completed parts of the cross-section and the new information (see example A.4 in the Appendices). This semiautomatic stage does not require any new additional input. The new information added in this stage by the user is a decision. If the new decision does not result in a unique outcome for that zone, or if the user wants to determine the points related to the unit contacts manually, the zone will be passed to the third (manual) stage (C.2 in the Appendices). Here, Geopropy can accept more geospatial information with the purpose of either singling out the outcome or accounting for specific manual considerations such as adding a contact point in a space between boreholes that is not based on the data in the database. Different options are available in the third stage: 1) connect existing points using the point IDs; 2) define new contact points and connect them to other new points or an existing point; 3) connect points in the same borehole from the left or right sides (which is normally useful if there is a change in the polarity of the unit in a borehole due to folds, for example);

and 4) skip making a decision for some points and let the algorithm complete the critical zone automatically. The new information added in the manual stage by the user can be geospatial information, a group of decisions and/or a determination that the algorithm has to complete the third stage by itself.

Finally, Geopropy completes the cross-section respecting all the geological units. To evaluate the contacts, the resulting geological areas in the cross-section will be compared to the raw data to determine whether all the raw data points are correctly situated in their respective zones or whether there are areas containing more than one geological unit. Moreover, if there is an area in which the geological unit is not identified by Geopropy, it will be flagged for further analysis by the user. At this point, a 3D visualization of the preliminary result will be displayed. An example of this visualization can be found in section A.1 in the Appendices.

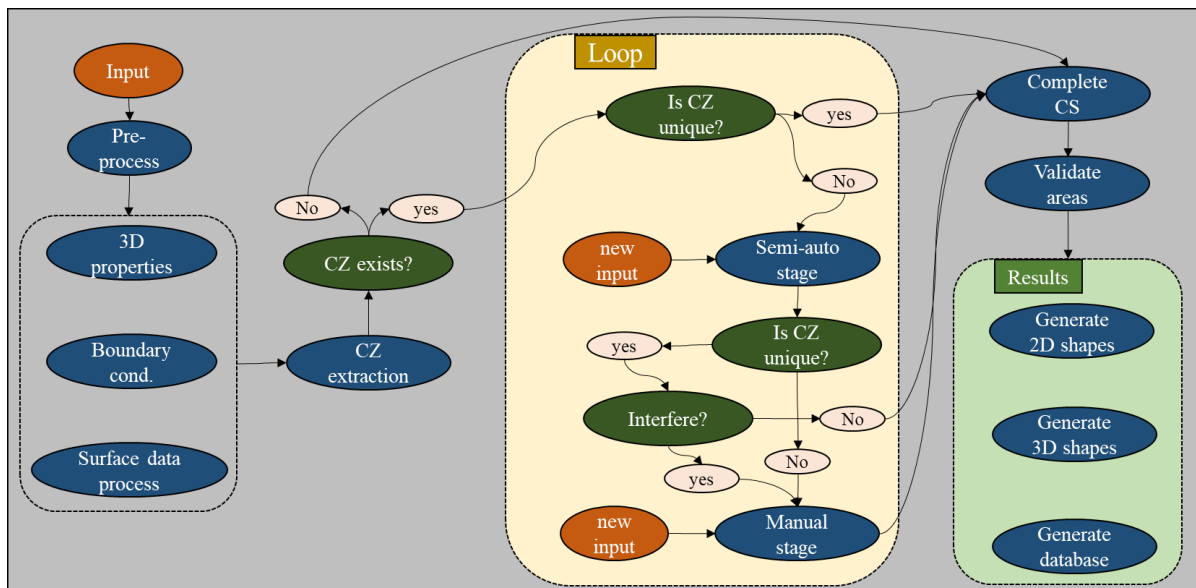


Figure 3 Simplified schematic Geopropy decision-making algorithm. CZ: Critical Zone. CS: Cross-section.

Some stages of the algorithm are described by a simple synthetic cross-section. Fig. 4 demonstrates the steps that Geopropy follows to generate it and the results. This cross-section contains conformable geological unit contacts, an unconformity contact and a normal fault, as shown in the coloured bars in Fig. 4.1. Tables B.1 to B.4 in the Appendices show the input data. This database is designed to visualize the simple decision-making process of Geopropy with a limited number of geological layers and boreholes.

After preprocessing the input data, the algorithm extracts the unit contact points (coloured dots in Figs. 4.2 and 4.3) based on the chronological data table (table B.3 in the Appendices). Then, it calculates the 3D properties of the cross-section, defines the boundary conditions and processes the ground surface data. The

side boundaries of the cross-section are the first- and last-introduced boreholes. The top boundary is defined based on the borehole and ground surface data. The preliminary bottom boundary is defined by a specific ratio of the maximum depth of the deepest borehole. The dark blue and violet rectangles in Fig. 4.2 show the top and preliminary bottom boundary points, respectively. Note that after creating all unit contacts, the bottom boundary will be modified by finding the maximum depth of the critical points and units between each pair of consecutive boreholes and reducing the area of the units at the bottom to decrease the areas without available information. The difference between the preliminary and final bottom boundaries is shown by the violet line in Fig. 4.3 and the cross-sectional bottom boundary line in Fig. 4. This cross-section does not contain any critical zones, so the algorithm directly passes to completing the cross-section automatically. Fig. 4.3 shows how Geopropy completes the cross-section in 2 steps: first, determining the unit contacts that have to be connected because they have just one other equal contact point on each side (green lines); then, determining the yellow contact lines, which the algorithm decides how to generate based on the relative age of the layers, orientation data and other information introduced in the database. After the completion of the cross-section, each geological body is determined in the form of a polygon or multipart polygons. Each colour and polygon name determines a specific geological material. The resulting geological bodies (polygons) are validated by comparing the polygons with the input geological unit data that are inside the polygon. If the resulting geological bodies cannot be validated by the input data, they will be flagged so that the user can analyse them in postprocessing. Finally, the outputs will be generated as shown in Fig. 4.4.

Although the top view of the boreholes does not have to be on a straight line, the general orientation of the boreholes has to follow a line in such a way that the same apparent dip angle for a contact through the whole cross-section results in negligible error. This is because for each unit contact, the apparent dip angle does not change between borehole locations since Geopropy draws a line between the holes.

Based on the needs, the user is able to ask for desired output shape files. The available options are unit contact lines, 3D vertical polygons for each geological unit, a 2D projection of the 3D cross-section and a curved version of the unit contact lines.

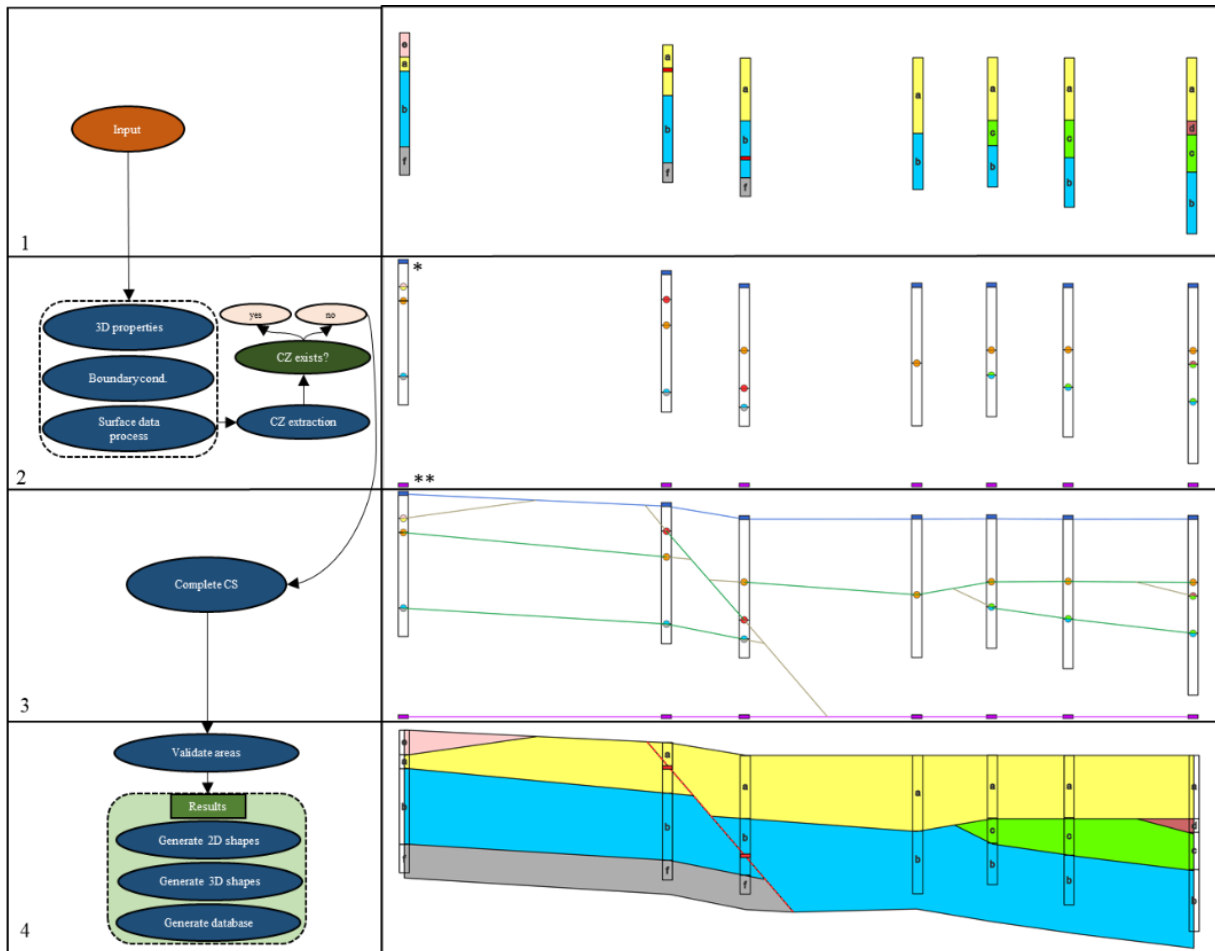


Figure 4 Geopropy algorithm steps used to generate a synthetic cross-section (CS). Red dots: fault points. Orange dots: conformable contacts. Two coloured dots: unconformable contacts. 1: Borehole input data 2: Contact point determination and critical zone (CZ) analysis. \* and \*\* show examples of the dark blue and violet rectangles that are used to determine the top and bottom preliminary boundaries, respectively. 3: Automatic contact line generation. Geopropy first generates green and then brown unit contacts. 4: Cross-section finalization. Polygons of the same colour identify a specific geological material.

To facilitate Geopropy usage, one main method (function, in the Python language) is designed to run the tool. The main method is executed with just two mandatory parameters: the input HYDOR database directory (which stores the geological information) and the desired borehole IDs. The other variables are optional: i) precision-related adjustments; ii) user preferences; and *iii*) visualization preferences. The developer mode allows the user a more powerful error handling option. Refer to section 2.6.1 for access to the library help for information about the method arguments.

### **2.3.4 Output and visualization**

For each critical zone in the semiautomatic stage, there is a 3D visual guide to illustrate the information graphically. Geopropy shows the final cross-section in a 3D environment. These two sets of graphs are based on the Matplotlib Python library <sup>109</sup>. Refer to Appendix A for some examples.

Various visualization options are available on demand. For each cross-section, in addition to 3D shape files containing interpreted geological units, it is possible to generate 2D projections of the cross-section. This is beneficial for illustration purposes, especially in the absence of a computer program that can visualize 3D shape files. Each visualization option generates a database that contains geological unit polygons as shape files, unit contacts as polylines and key coordinates as points. Depending on the visualization platform in use, it is also possible to modify the graphical details, such as the colour, boundaries and line type. Additionally, there is a built-in option to smooth the geological unit contacts using curved lines on demand. This option is designed for use in 2D and is only an aesthetic tool. Smoothed cross-sections are not recommended since they result in decreased model accuracy, which could cause potential difficulties in the subsequent processing stages, such as fence diagrams and 3D volumes. The results can be displayed in any computer program that supports 3D shape files, such as ArcGIS or QGIS. Although there are certain shortcomings in 3D data processing and visualization (“Editing polygons in 3D—ArcMap | Documentation”), we used the Arcpy Python library to configure and optimize the output shape files.

## **2.4 APPLICATION TO SYNTHETIC DATASETS**

Three different synthetic datasets with varied properties are designed in 2D for better visualization. Each dataset was input to Geopropy and given to a geologist to generate the cross-section. The geologist used HEROS to visualize and create the cross-sections. The first dataset is described in section 2.3.3 to demonstrate the functionality of the algorithm. In this database, there are geological units that do not exist in all boreholes, so they have to terminate according to the guidelines and assumptions. The second synthetic dataset contains conformities, unconformities, normal faults and an intrusion. The third dataset contains an intrusion, a reverse fault in conformable contacts and another reverse fault that crosses through a fold. This dataset focuses on testing the semiautomatic and manual stages of Geopropy when there is more than one possible correct explanation for the same dataset in each zone, whether it is an intrusion zone or fold-fault zone.

## 2.4.1 Synthetic database 1

### Dataset overview and input

The database directory and borehole IDs that are mandatory arguments to execute Geopropy correspond to Tables B.1 to B.4 in the Appendices, which describe the geological specifications of the example. It is assumed that the only available data in this example correspond to boreholes, that is, there are no ground surface data available.

### Results and comparison

Fig. 5 shows the cross-sections completed by a geologist and Geopropy. For the most part, the two cross-sections are similar, although there are some differences:

- 1) Units  $e$  and  $d$  have different forms. The termination of a unit contact in the absence of information can be done in various ways. In this case, the geologist chooses a smoother line, whereas Geopropy draws a straight line into the middle of two boreholes.
- 2) Similarly, the intersection of units  $a$ ,  $b$  and  $c$  (red boxes in Fig. 5) is slightly different and again is not smoothed by Geopropy.
- 3) Layer  $f$ , which appears only on the boreholes on the left side of the fault, is not continued by Geopropy on the footwall since there is no information in the database to verify that this layer continues on the right side of the fault, whereas it is drawn there by the geologist (Fig. 5.1).
- 4) Regarding the depth of the cross-sections, Geopropy has diverse ways of interpreting the deepest boundary of the cross-sections. Refer to section 2.6.1 for access to the Geopropy documentation.

The first two differences can be considered as simply dissimilarities in styles.

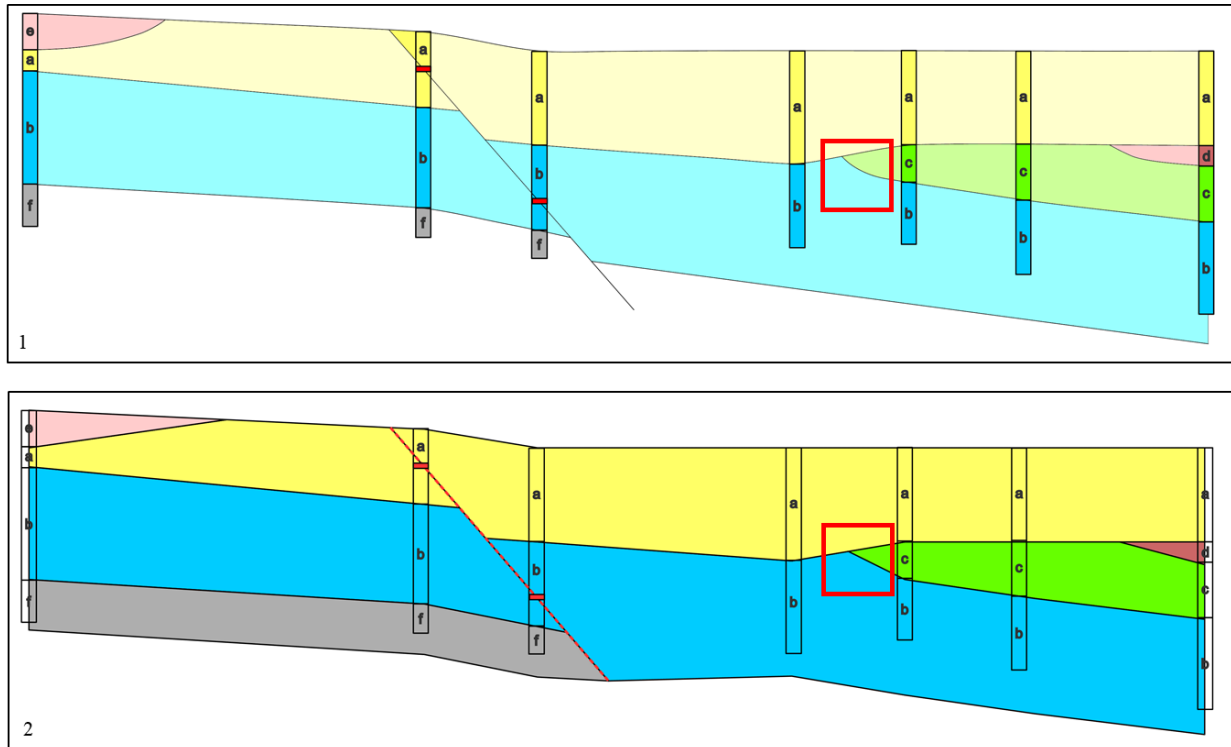


Figure 5 First synthetic example. The colours and letters correspond to geological units. The example contains a normal fault. 1: The raw data are illustrated in the coloured columns, and the cross-section was completed by the geologist. 2: Cross-section generated by Geopropy with raw data projected.

## 2.4.2 Synthetic database 2

### Dataset overview and input

This dataset is inspired by Earth Science student exercises. It includes conformable and unconformable contacts, with older faults and an intrusion of relatively complicated geometry. The coloured bars in Fig. 6.1 illustrate the raw borehole data. Tables B.5 to B.8 in the Appendices show the data used for this example.

### Results and comparison

The cross-sections generated by the geologist and Geopropy are compared in Fig. 6. Similar to the first example, the general geometry of the two cross-sections is very similar. The intrusion (unit *a*, in yellow in Fig. 6) contacts are smoother in the geologist's cross-section than in Geopropy's cross-section. In addition, in the two contact zones between the intrusion and unit *n* (coloured dark blue in Fig. 6), the geologist softened it to draw a zone contact, in contrast to the singular contact points in Geopropy. This is the major difference between the two, and it could also be considered a personal, style-related decision.

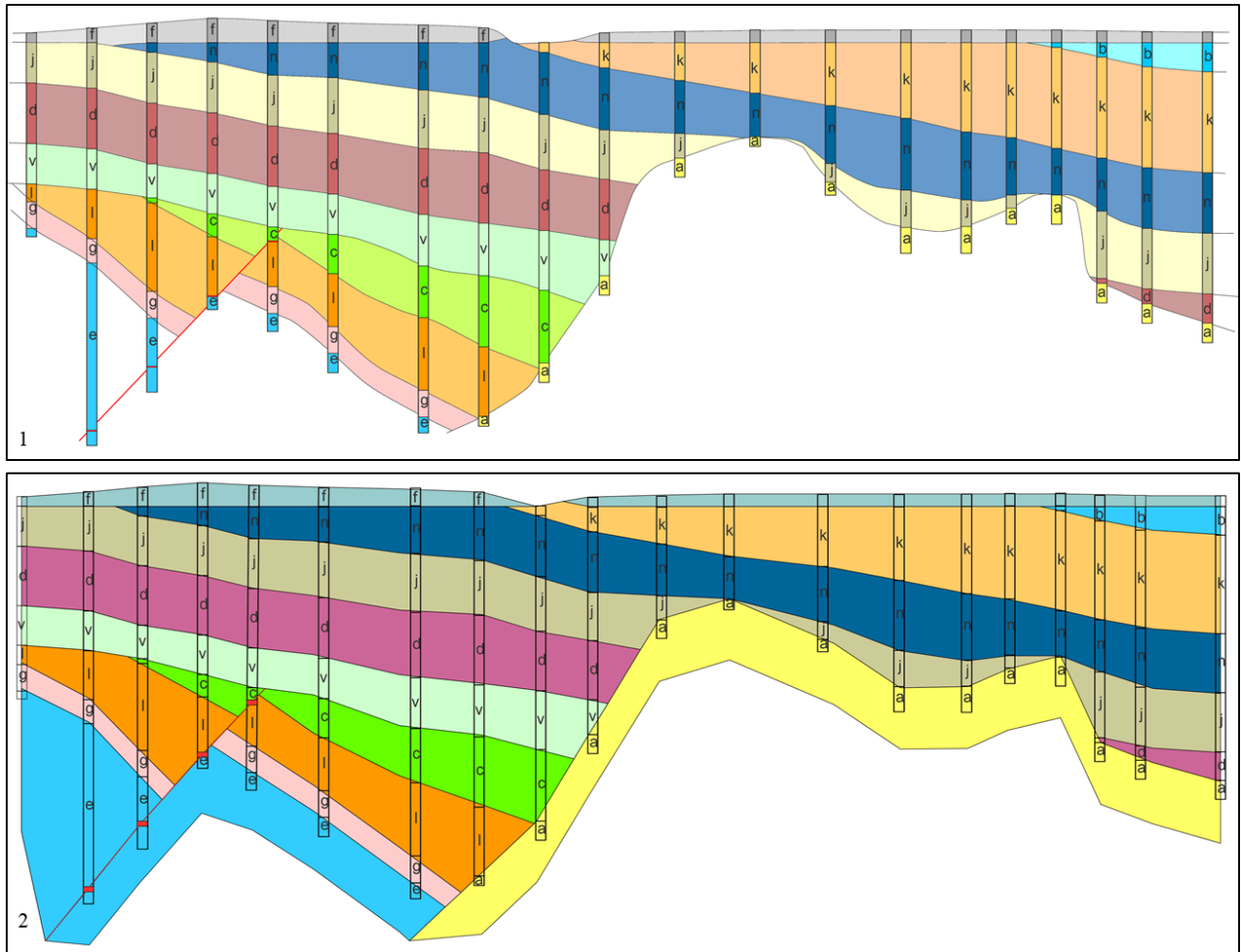


Figure 6 Second synthetic example. The colours and letters correspond to geological units. This includes a normal fault and an intrusion. 1: The raw data are illustrated in the coloured columns and the cross-section completed by the geologist. 2: Cross-section generated by Geopropy.

### 2.4.3 Synthetic database 3

#### Dataset overview and input

The goal of this example is to challenge Geopropy with scenarios with critical zones that can lead to several possible outcomes. Fig. 7 illustrates borehole and ground surface data, which in turn contain two sets of ground surface information. At some points, only the altitude is available (black points), whereas at others, there is information about the unit contact (Fig. 7, zone *i*, coloured points). Each point indicates a unit contact, and each colour indicates the unit that has been observed.

Zone *ii* includes a reverse fault that causes repetition of units *a*, *b* and *c* in all boreholes of the zone. Zone *iii* is a region with complex geometry that could be interpreted in different ways. There must be folds and a reverse fault (Figs. 8.1 and 8.2). Zone *iv* illustrates an intrusion that can have more than one possible geometry. Refer to section 2.6.2 for the details of how to reproduce the example.



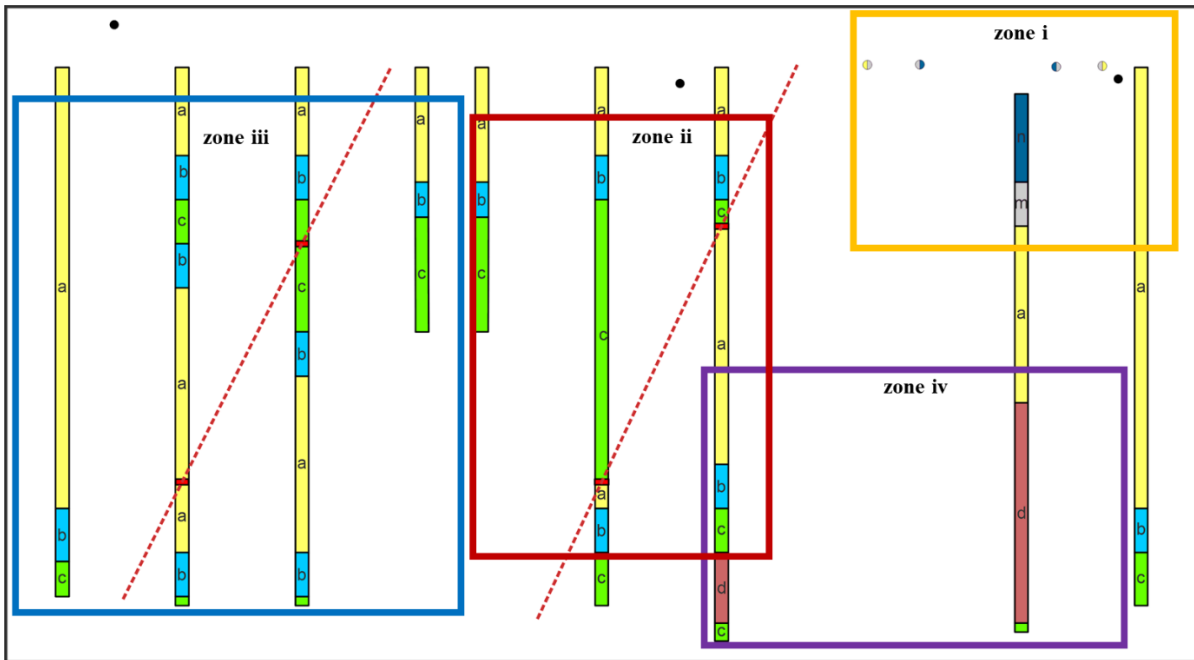


Figure 7 Raw data of the third synthetic example. The circles show the available surface data. The coloured circles show information about the geological units. The zone contains 2 reverse faults (zones *ii* and *iii*), an intrusion and a fold.

Geopropy detects that the dataset does not result in a unique outcome immediately after introducing the dataset directory and the borehole IDs. Thus, it starts an interaction with the user asking for complementary information. This interaction is discussed below. Tables B.9 to B.13 in the Appendices show the tables in the dataset. List 1 shows how to execute the function.

```

1. #Import GEOPROPY library
2. import geopropy as gpp
3. #Execute main function
4. lit_table="Borehole_subunits"
5. bore_ids=[111,112,113,114,115,116,117,118,119]
6. database_dir="database directory"
7. gpp.cross_section(database_dir, bore_ids, lit_table)

```

List 1 Executing synthetic dataset 3 using the GEOPROPY library. Note that the default arguments of the cross\_section method are used. For more information, refer to section 2.6.1.

## Results and comparison

The interpretations of zone *i* by the geologist and Geopropy are fairly similar (Fig. 9). The Geopropy contacts are more angular, whereas the geologist's contacts are more rounded or smoothed, as in the previous examples. In zone *ii* (Fig. 7), where units *a*, *b* and *c* are repeated in the boreholes, the existence of

the reverse fault shapes the geometry of the zone. Therefore, although Geopropy identifies this region as a critical zone, it does not ask for more information since the outcome is unique (Fig. 9).

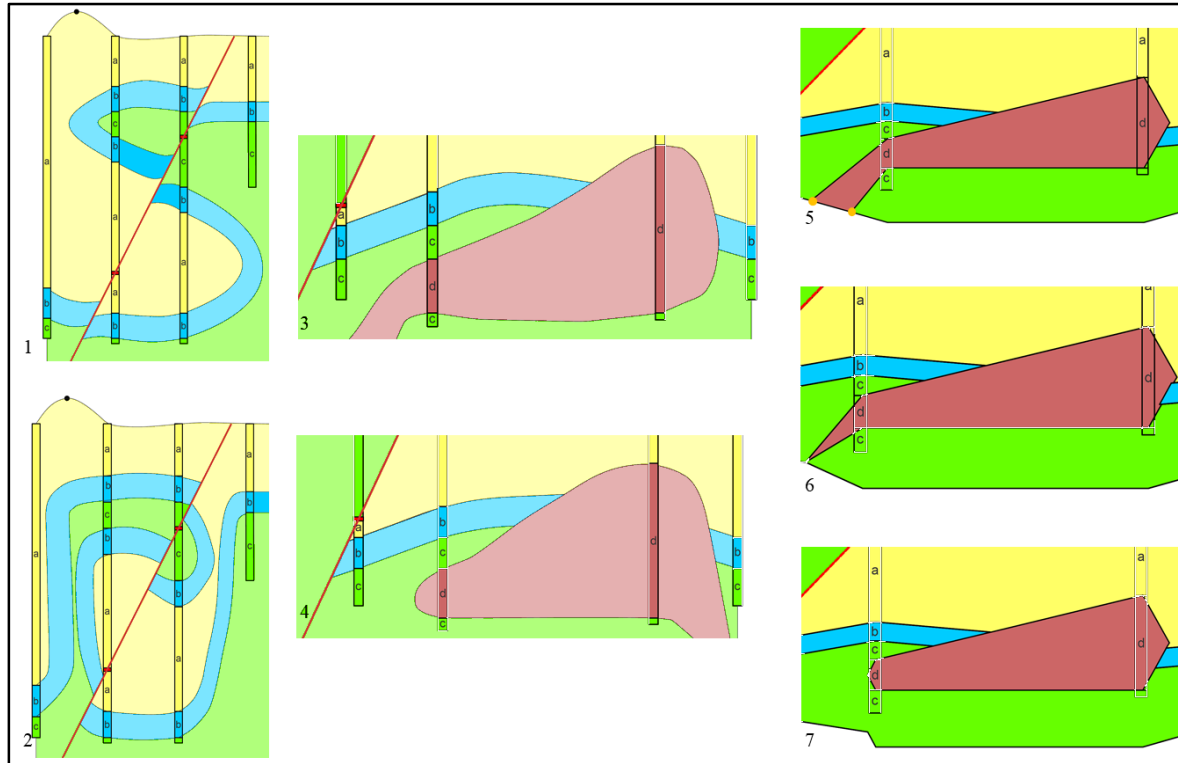


Figure 8 1 and 2: Two possible interpretations of zone *iii* drawn by the geologist. 3 and 4: Two possible interpretations of zone *iv* drawn by the geologist. 5: Zone *iv*: Geopropy results in the manual stage by the choice of the user, taking into account the same assumptions as 3. The orange dots were added manually by the user. 6 and 7: Geopropy interpretation of zone *iv*, completed in the semiautomatic (guided) stage.

The results obtained in critical zone *iii* (Fig. 7) contain repetitions of the geological units, but the unit contacts can be interpreted in more than one way. In addition, there is a reverse fault crossing this critical region. Figs. 8.1 and 8.2 show two of the possible outcomes, which were drawn by the geologist. To reach a unique solution in this critical zone, Geopropy proceeds to the semiautomatic stage, where the user has to choose one of the possible scenarios based on preference or experience (listing C.2 in the Appendices ). Fig. A.4 in the Appendices shows the visual 3D helper of Geopropy used to facilitate the procedure. Since the information provided by the user in the semiautomatic stage results in a unique solution in the critical zone, Geopropy verifies the uniqueness of the critical zone and does not enter the manual stage. Fig. 8.1 shows the geometry of the units generated by the geologist in critical zone *iii*, which is also introduced to Geopropy. The comparison between Figs. 9.1 and 9.2 in zone *iii* illustrates the difference between them, which is the lack of smoothness, especially for the contact points between unit *b* and the fault and respecting the thickness of unit *b* in the geologist's interpretation.

The intrusion in critical zone *iv* could be interpreted in various forms (Figs. 8.3 and 8.4), according to the geologist. Geopropy detects the nonuniqueness of the zone and proceeds to the semiautomatic stage to ask for more information from the user, which could lead to different cross-sectional outcomes based on different choices of the user, as shown in Figs. 8.5, 8.6 and 8.7. Figs. 8.6 and 8.7 are generated in the semiautomatic stage, whereas Fig. 8.5 is generated in the manual stage. In addition, it is possible to introduce new sample points to the zone using spatial data, depending on the user needs and available data. Listing C.3 in the Appendices shows the Geopropy user interaction in the semiautomatic and manual stages.

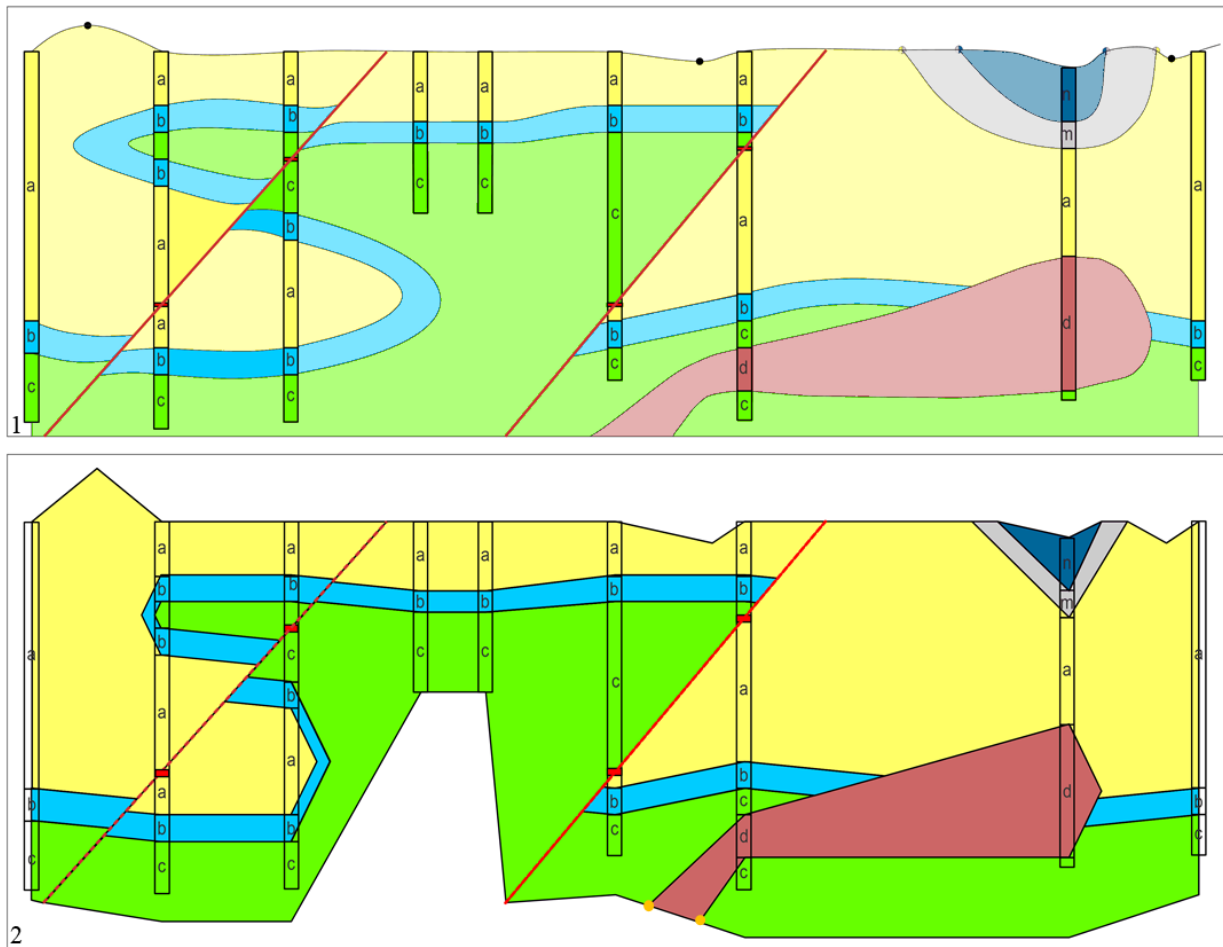


Figure 9 Results of the third synthetic example. The colours and letters correspond to geological units. 1: Raw data illustrated in the coloured columns and the cross-section completed by the geologist. 2: Cross-section generated by Geopropy with the same settings as 1. Zone *iii* is completed in the semiautomatic (guided) stage, and zone *iv* is completed in the manual stage based on the choices of the user. The orange dots are added manually by the user.

There are some differences between the interpretations of the geologist (Fig. 9.1) and Geopropy (Fig. 9.2). Smoothness is the first difference, especially around folds and intrusions. The unit contact angles also differ. In the absence of angular data, the geologist relies on experience and personal style, whereas Geopropy calculates the apparent dip in the unit contacts based on the contact lines across the whole cross-section to

preserve the orientation behaviour of the unit. It must be mentioned that identifying the available options and being able to apply additional information offers freedom to the user to obtain results that fit expectations. Figs. A.3 and A.5 in the Appendices illustrate two other possible outcomes created by the geologist and Geopropy, respectively.

## **2.5 DISCUSSION AND CONCLUSION**

Geopropy is an open-source Python library that performs as an intelligent agent to create 3D geological cross-sections based on a knowledge- and data-driven approach. As seen in the three presented examples (Figs. 5, 6 and 9), Geopropy's cross-sections preserve the overall characteristics of a real geologist's cross-sections, although small differences, mostly in the smoothness of contacts, can be observed. The application of Geopropy to synthetic profiles validates the functionality and the decision-making procedure, making it a useful support tool for geologists.

The use of Geopropy outputs in a GIS platform may aid users in modifying and customizing the results with the capabilities available on each GIS platform. For instance, using a GIS, fence diagrams or geological volumes can be easily generated, which can be used later in numerical modelling software.

Nevertheless, Geopropy has some limitations. First, it is highly sensitive to the data. Geopropy always respects the input data, whereas in some scenarios, a geologist may decide that an observation is not reliable or not compatible with the properties of the region and may ignore or modify the data. In Geopropy, this step has to be carried out by the user. Second, although Geopropy is an open-source library, it is not completely free to use since it depends partially on the Arcpy library to generate the shape files. This means that to execute Geopropy, the user needs to have access to an ArcGIS licence. The reason that other freely available libraries were not used to create shape files is that there was no free and open-source library that could reliably support 3D shape file generation. Even with the shortcomings of the Arcpy library, we considered it the most suitable library for Geopropy. In addition, there is a vast range of tools developed and available in ArcGIS that could be coupled with Geopropy or used in the postprocessing of the results. Third, in the case of high variation of the angles among the vertical planes between each consecutive borehole pair in a cross-section, the assumption of the same apparent dip angles for one contact along the whole cross-section in 3D could result in considerable errors. Fourth, Geopropy does not necessarily preserve the thickness of the geological units. It is assumed that by preserving the unit contacts and the unit angles, the output unit thickness will be accurate, but thickness is not considered an ascertainable factor in generating the cross-sections.

Developing Geopropy in a widely used programming language and making the source code available create the possibility for further development of the library and coupling the program with other existing decision-making tools. In addition, the use of a standard and well-known database, with various tools depending on it, facilitates the integration of different aspects of the study and saves the time needed to prepare databases for each tool. Moreover, by following specific guidelines that are derived from the thought process of a geologist, i) the process of generating the cross-section and the results can be easily interpreted since the algorithm acts similarly to the user's thinking steps; ii) the code controls inconsistencies related to personal style and bias, which are important in studies that involve more than one person or that continue for a long period; and iii) the code can help inexperienced users avoid the unrealistic results that could occur when complicated mathematical modelling techniques are used.

Geopropy is easy to implement, and keeping in mind that creating large numbers of cross-sections is time consuming, it could be used to speed up the process of creating geological cross-sections. Overall, this tool could be of great help when geological modelling must be done explicitly, as it can avoid inconsistent decision making. In addition to the time benefits and avoidance of inconsistencies, another valuable asset of Geopropy is that it detects zones with several possible outcomes in a cross-section. This could potentially help users analyse different plausible geological scenarios based on the available data. This also holds in the absence of orientation data, which brings flexibility to the tool at the cost of precision. Geopropy does not replace implicit models, but it would help in generating cross-sections explicitly as an intelligent agent.

## **2.6 SOFTWARE AND DATA AVAILABILITY**

### **2.6.1 GEOPROPY library information:**

Available to download freely in <https://github.com/IDAEA-EVS/Geopropy> Under AGPL-3.0 License.

Year first available: 2021

Dependencies: Arcpy (ArcGIS 10.5 or higher), Matplotlib, pypyodbc

Programming language: Python

Developed by Ashkan Hassanzadeh.

Contact information: [ashkan.hassanzadeh@csic.es](mailto:ashkan.hassanzadeh@csic.es)

Refer to <https://github.com/IDAEA-EVS/Geopropy/wiki> for additional information about the installation, default values of the arguments, the explanation and the usage.

## 2.6.2 Synthetic examples

Databases of 3 synthetic examples in this article are available free of charge in .mdb format alongside the Jupyter notebook in <https://github.com/IDAEA-EVS/Geopropy>

### **3 An open source Python library for environmental isotopic modelling**

This chapter is based on:

**Hassanzadeh, A.,** Valdivielso, S., Vázquez-Suñé, E., Criollo, R. & Corbella, M. An open source Python library for environmental isotopic modelling. *Scientific Reports* 2023 13:13, 1–19 (2023).  
DOI: [10.1038/s41598-023-29073-2](https://doi.org/10.1038/s41598-023-29073-2).

### 3.1 INTRODUCTION

Water isotopic composition is of paramount importance for decision making in many fields of study, including environmental resource management <sup>111</sup>. The stable water isotopes <sup>18</sup>O and <sup>2</sup>H are indicators of diverse aspects of the hydrological cycle.  $\delta^{18}\text{O}$  and  $\delta^2\text{H}$  measurements in precipitation are utilized in different meteorological and hydrological studies to identify the origin of precipitation, recognize local effects in water cycle studies, define the relative shares of water with different origins in a water body, describe aquifer recharging and characterization process and investigate various aspects of runoff and stream flow generation. All these features are essential for the optimal and sustainable management of water resources <sup>112,113</sup>.

The isotopic composition of rainwater is influenced by different physical variables and processes: temperature; pressure; humidity during condensation (to generate precipitation) <sup>114,115</sup>; mixtures of air masses with distinct origins <sup>116</sup>; the isotopic composition of the seawater from which air moisture condenses <sup>117</sup>; in-cloud microphysical processes <sup>118–122</sup>; the moisture conditions below clouds and the partial evaporation of precipitation along the path between clouds and the ground <sup>123–125</sup>; and the mixture of recycled precipitation from evapotranspiration over continents <sup>126–128</sup>. Therefore, detailed isotopic signature studies are used to discern these effects in any study area.

A linear relationship called the global meteoric water line (GMWL) is present between the  $\delta^{18}\text{O}$  and  $\delta^2\text{H}$  of meteoric water at the global scale, and this relationship is defined as  $\delta^{18}\text{O} = 8 * \delta^2\text{H} + 10$  <sup>124</sup>. The characteristic isotopic signature of meteoric water in a particular region is caused by the various temperatures, relative humidity values, amounts of precipitation, latitudes and landmass proximities. The water molecules components (O, H) undergo isotope fractionation during phase transitions and the ratios of heavy versus light isotopes acts as a traceable feature of the physical processes <sup>129–133</sup>.

Two common approaches are available for studying the global distribution of the isotopic composition of precipitation: isotope-enabled atmospheric general circulation models (IGCMs) and regression statistics-based approaches <sup>17</sup>. IGCMs are numerical models that improve our understanding and reveal valuable information of the atmosphere by considering different physical processes (diffusion, advection, convection, etc.), including the physics of water isotopes (e.g., isotope fractionation, evaporation, condensation, among others) <sup>134</sup>. Computational power and numerical modelling advancements in recent decades have played an important role in the development of IGCMs, as they have resulted in a variety of models at different regional scales with diverse levels of complexity, such as CAM5 <sup>135–137</sup>, ECHAM5 <sup>138,139</sup>, MIROC <sup>140</sup> and LMDZ4 <sup>141</sup>. IGCMs are usually complex, time consuming and computationally demanding simulations. On the other hand, regression statistics-based models are generally useful in identifying the possible processes suffered by water samples based on their isotopic signature. Statistical models are simple



to apply and are more intuitive to interpret. Consequently, they are used as stand-alone or complementary – preliminary tools for interpreting IGCM models and evaluating their results <sup>134</sup>.

Statistical models exhibit some shortcomings that can limit their usage or lower their precision. First, in contrast with IGCMs, there is no specific standalone tool that allows the user to determine the input features and databases for developing a statistical isotopic model. Second, some study areas possess scarce isotopic data or different types of isotopic samples (individual rain events versus accumulated events) and/or contain meteorological measurements with diverse spatiotemporal resolutions. This may limit the usage of the available variables that can affect statistical isotopic models <sup>17,19</sup>. Third, most statistical regression studies are based on simple linear models, which can neglect some of the underlying processes of the water isotopic signature by not exploring the more complex relationship between the variables. The use of both standard and novel mathematical approaches can explore these possibilities and could potentially result in discovering unforeseen aspects <sup>142</sup>. Fourth, the use of statistical analyses can be time- and effort-consuming, depending on the type and number of models needed or the output desired (meteoric water lines, estimation graphs, detailed maps, etc.). Automatically creating an extensive output could prevent systematic errors without compromising the possibility to carefully examine the significance and relevance of the inputs and results by the user, if it is accompanied by the informative reports of each underlying processes.

To address these shortcomings, we present Isocompy, an open source, Python-based, multistage isotopic composition analysis and modelling library. The main objectives of Isocompy are (i) to introduce an open source framework that integrates the diverse steps of stable statistical isotope modelling in a dedicated library; (ii) to incorporate novel data management, statistical analysis and machine learning regression methods accompanied by decision-making algorithms; (iii) to exhibit flexibility regarding the available input data and function with measurements that are scarce and discontinuous in time and heterogeneous in space.; (iv) to be intuitive and user friendly, which speeds up the process of forming an isotope model; and (v) to generate reports and figures in every step if needed so that the user can understand the ongoing procedure.

In the following sections, we describe the methods used (section 3.2) and the different aspects of Isocompy (section 3.3), and we demonstrate its functionality by applying it to an example involving Salar de Atacama (Chile) (section 3.4).

## **3.2 METHODS**

To create the Isocompy algorithm, bibliographical research is performed to define the innovative capabilities that would be needed for isotopic modelling. The workflow of the program is then chosen accordingly. In this section, we discuss the necessity of the capabilities that are included in Isocompy and

the methodology used in the proposed workflow to form the isotopic precipitation composition models with respect to the aforementioned objectives.

Various input parameters can affect the isotopic composition of rainwater. Meteorological (precipitation, relative humidity, temperature, etc.) and geospatial parameters are two groups of input data that are widely used in isotopic modelling<sup>143–149</sup>. However, other information may be needed, such as sea surface temperatures, atmospheric pressures, outgoing longwave radiation (OLR) values<sup>120,150–153</sup>, features derived from air mass trajectories<sup>154,155</sup> or features resulting from reanalysis (such as wind components, dewpoint temperatures, and evaporation values).<sup>156</sup>. Therefore, the workflow must allow the user to choose the nature of the input features. Moreover, in cases where the database contains unwanted data for the ongoing study, it can be modified easily.

Furthermore, some industrial and scientific projects are carried out in regions with limited or discontinuous spatiotemporal data. For example, in some cases, the meteorological stations are continuously maintained, which results in the production of a long-term dataset. Conversely, isotopic measurements are often sparse in time and poorly distributed in space and are not necessarily measured at the same position as other input parameters (e.g., weather parameters); this is mostly due to the complexity and costs of the analyses. Fig. 1.1 illustrates an imaginary example of two independent parameters (red crosses and blue circles) that potentially affect the isotopic measurements (green triangles), but since they are not available at the same location, a one-to-one relation between the features cannot be made to perform regression. Moreover, the densities of the available data are different among the red crosses, blue circles and green triangles. To obtain of the features at the green triangle positions, first, regression models for the red and blue points, which are variables dependent on other features (in this case, geospatial features), must be generated. Figs. 1.2 and 1.3 illustrate the estimations of each red and blue feature obtained at the green triangle positions derived from two separate regression models (F1 and F2, respectively).

In the yellow diamonds in Fig. 1.4, the calculated values of the red and blue parameters (estimated from F1 and F2) and the corresponding green measurements are available, which makes it possible to construct a regression model with red and blue parameters as independent variables and the green parameter as a dependent variable. By using this model, it is possible to create a map of the green parameter (isotopic composition).

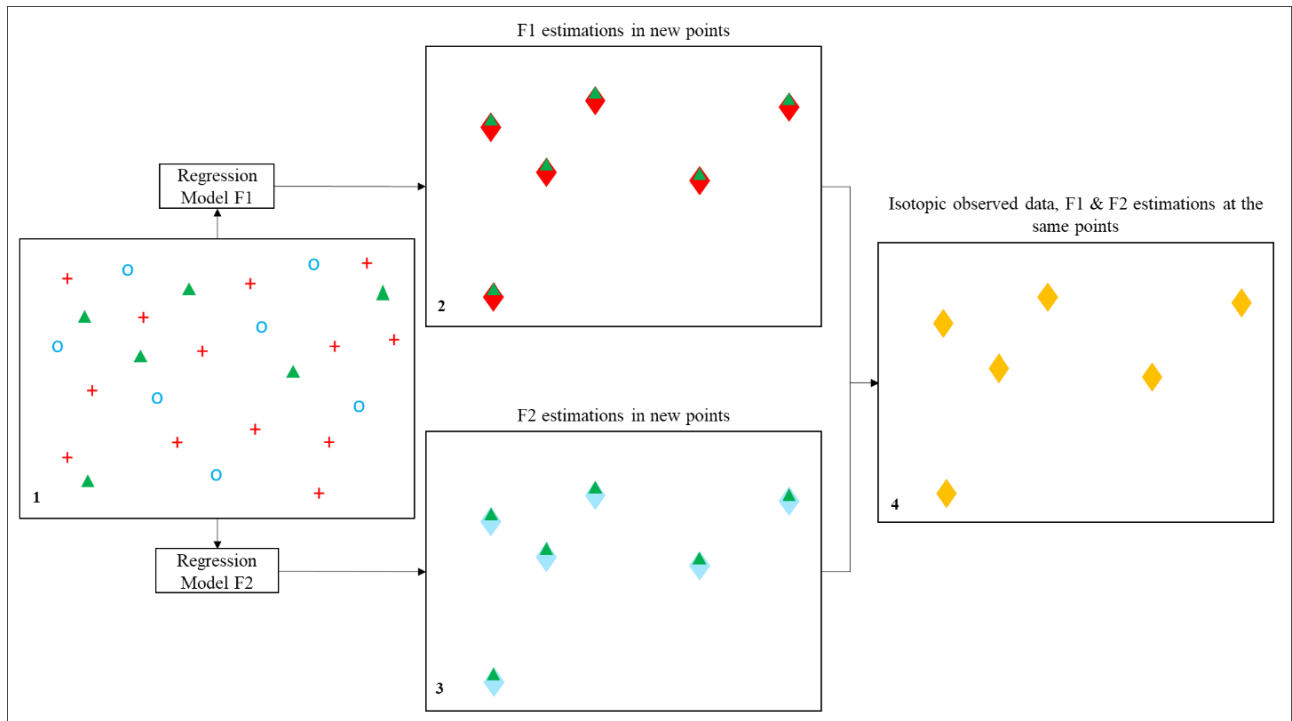


Figure 10. Workflow scheme for estimating isotopic values by using two independent parameters that are available in different locations than the isotopic measurements. Blue circles and red crosses represent two independent features that are potential candidates affecting the green triangles (isotopic precipitation composition). 1) Imaginary map of all available points. Green triangles are points with field isotopic measurements, and red crosses and blue circles are two independent parameter measurements. 2 and 3) Estimation of the red and blue parameters from the constructed F1 and F2 regression models at the location points where isotopic data are available. 4) As a result of the algorithms, in the yellow diamonds, estimated red and blue data and measured green (isotopic composition) data are available.

Data can also vary within a time window. Utilizing the example in Fig. 1, let us assume that a continuous and dense amount of data are available for the blue parameter in a specific month during ten consecutive years. However, in the red crosses, data are available in the same month for six years. To overcome the limitations derived from different measurement frequencies and time windows, one solution is to average the ten- and six-year measurements in the blue and red points, respectively, to obtain a single set of data for each feature in each location. Although averaging the measurements may result in a loss of information while producing less precision and higher model uncertainty, these effects would also occur with other data treatment techniques, such as filling the gaps in data series. The final workflow must also account for different parameters that are measured directly alongside the isotopic water composition or estimated via other methods, such as features derived from reanalysis<sup>157,158</sup>. Another important aspect of the workflow is to analyse the degrees of influence of suspected features on the dependent variable. Considering that the goal is to produce a workflow that is simple yet precise, an automatic statistical analysis procedure based on multicollinearity examination and a feature selection algorithm must be crucial parts of the workflow.

The fact that the relations between earth science variables may be linear or nonlinear suggests the capability to apply different regression methods in the workflow. The regressions must be accompanied by calibration and validation procedures to find the regression method with the highest estimation power that avoids common modelling errors such as overfitting. A total of eight regression models are considered in this study: Elasticnet <sup>159</sup>, Bayesian ridge regression <sup>160</sup>, least-angle regression <sup>161</sup>, Bayesian automatic relevance determination (ARD) <sup>162</sup> and orthogonal matching pursuit <sup>163</sup>, support vector regression <sup>164</sup>, a random forest <sup>165</sup>, and a multilayer perceptron <sup>166</sup>. Since some of these methods are sensitive to the data scale, all inputs are standardized before applying the regressions. Hyperparameters are parameters of machine learning methods whose values control the learning process <sup>167</sup>. The brute-force hyperparameter search algorithm is used to obtain a suitable set of hyperparameters <sup>168</sup>; it is optional to fit regression methods to the transformed  $\ln(1+x)$  of the input data alongside the original data which can potentially result in a better model in case the features have log-normal distributions.

In Elasticnet, both L1 and L2 regularization terms are used to avoid overfitting. The Lasso (L1) and ridge (L2) regression methods are specific forms of Elasticnet regression, where the former adds the absolute value of the magnitude and the latter adds the squared magnitude as a regularization term to the cost function. Lasso and ridge regressions are achieved by introducing an L2 to L1 ratios equal to zero and one, respectively. A more detailed description of this method can be found in <sup>159</sup>.

The orthogonal matching pursuit method constrains the number of zero coefficients. Its residuals are calculated by using an orthogonal n-dimensional projection, which assumes, similar to independent variables, that the dependent variable can contain measurement errors <sup>163</sup>.

Least-angle regression is a stepwise linear regression method that moves in the direction of the most correlated feature in each step. This method is beneficial when the number of features is higher than the number of samples. Least-angle regression is sensitive to outlier data <sup>161</sup>.

The Bayesian ridge and Bayesian automatic relevance determination methods (also known as sparse Bayesian learning and relevance vector machine regression, respectively) form probabilistic models that include regularization parameters that are tuned according to the available data instead of being defined prior to regression <sup>160</sup>.

Random forest regression is a method based on the average of randomized independent decision tree estimator outputs. The main concept of this method is that the integrated final estimator may produce better results than any of the single decision trees since combining them decreases the standard deviation of the estimates <sup>165</sup>.

Support vector machines are versatile supervised learning methods that are used in various environmental science fields <sup>169</sup>. They can be used in high-dimensional environments and are flexible depending on the chosen seed functions. It must be taken into account that support vector regression can be computationally demanding <sup>170</sup>. Moreover, if the number of features is higher than the number of samples, the seed functions must be selected in a way that avoids overfitting <sup>171</sup>.

Neural networks have proven to be effective estimation techniques in various branches of science. Multilayer perceptron regression is a supervised learning method that uses L2 regularization to avoid overfitting the weights. An MLP uses a backpropagation technique. The ability to determine the number of hidden layers, the size of each layer and diverse type of activation functions mark an MLP as a flexible technique <sup>166</sup>. However, an MLP is complex during the process of choosing the correct estimator hyperparameters.

### 3.3 UNDER THE HOOD OF ISOCOMPY

#### 3.3.1 Isocompy workflow

Considering the abovementioned aspects of isotopic composition modelling, Fig. 2 illustrates the general scheme of our proposed workflow. It consists of data preparation and two main stages. The independent variables are introduced in the data preparation step. The goal of the first stage is to estimate the independent parameters that affect the isotopic composition model in the same space-time framework as the empirical data. The results of the first stage, accompanied by the empirical data, are incorporated into the second stage to obtain  $\delta^{18}\text{O}$  and  $\delta^2\text{H}$  models. Stage one of the workflow begins with a statistical analysis of the independent variables that are introduced in the data preparation step to determine their degrees of influence on the dependent variable and select the substantial variables for the regression model. The regressions are applied, and the most calibrated model is selected. Then, the variables that influence the water isotopes are estimated in the same time and space as the isotopic measurements. By preparing the data from three source groups (estimated variable data, measured variable data and measured isotopic data, (1.4, c and b in Fig. 2, respectively)), it is possible to obtain isotopic models in stage two. Again, a statistical analysis leads to the extraction of the substantial independent variables over which the regressions will be applied to select the best model. Once the models are available, the isotopic composition values can be estimated. The underlying sections of each stage are explained below.

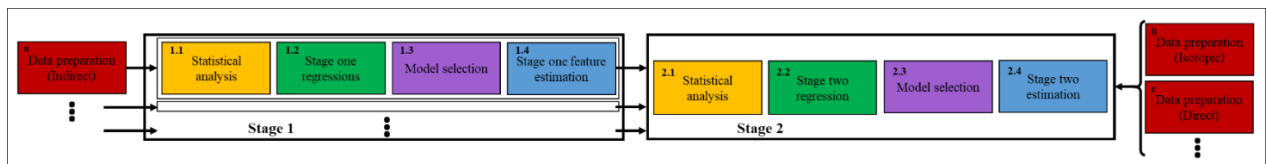


Figure 11. Scheme of the Isocompy workflow utilized to design the Isocompy architecture. It consists of data preparation (red boxes) and two main stages. Each stage includes statistical analysis (yellow boxes 1.1 and 2.1), regression (green boxes 1.2 and 2.2), model selection (violet boxes 1.3 and 2.3) and feature estimation steps (blue boxes 1.4 and 2.4).

Data preparation (the red a, b and c boxes in Fig. 2) is a key step that defines many major properties of the constructed model. Box a in Fig. 2 shows the input features named indirect features since they are not measured with isotopic values; box b represents the isotopic input measurements, and box c illustrates other features measured directly with isotopic values (direct features). In the data preparation step, different aspects of the model must be determined by the user.

- The dependent and independent variables.
- The temporal window of the input measurement choice.
- Input filtration based on specific time properties, if needed (El Niño or La Niña Southern Oscillation).
- Outlier removal based on diverse methods, if needed.
- The data averaging technique.
- Brute-force searching hyperparameter definition.

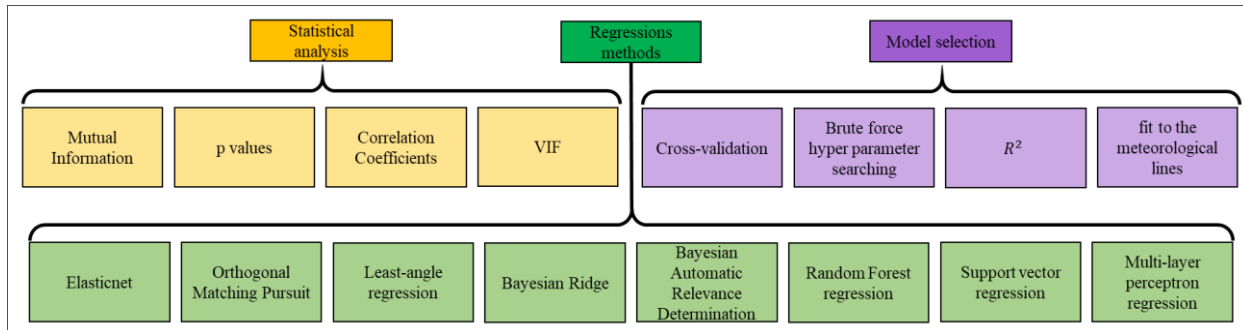


Figure 12. Yellow, green and violet boxes show the techniques used in the statistical analysis step, the regression methods available in Isocompy and the implemented techniques in the model selection steps, respectively.

The statistical analysis step (yellow boxes in Fig. 2) allows the algorithm to select the most considerable features to be used afterwards in the regression models. Feature selection is crucial in environmental models that normally use spatial features as inputs since autocorrelations in data may distort the estimation power of the model<sup>172</sup>. As shown in the yellow boxes in Fig. 3, in this stage, the algorithm calculates the p values determined by one-tailed F-test on centred data, mutual information<sup>173</sup>, correlation coefficients and variation inflation factors (VIFs) of the variables. The p values and mutual information help to determine the linear and nonlinear relationships between parameters and evaluate the significance of the parameters<sup>174,175</sup>. The VIFs and correlation coefficients are useful for detecting multicollinearity.

Since one of the main objectives of the algorithm is to facilitate and speed up the model generation process, the feature selection procedure that is derived from the statistical analysis can be performed automatically or controlled by user-defined or predefined values. In the automatic mode, the algorithm first uses the VIFs, correlation coefficients and optional pairs of features defined by the user to remove the features with multicollinearity effects that higher than a defined threshold. Then, p values are used to select the statistically significant features, based on the user-defined alpha level.

It is important to mention that since the F-test assumes that the features are distributed normally, the user have to check the normality of the features that are chosen as important features in VIF test.

The regression steps are performed in two stages of the algorithm (green boxes in Fig. 2). Various linear and nonlinear regression methods are available, as shown in the green boxes in Fig. 3, which can be selected by the user based on the nature of the given study or computational power, among other strategies. The regression methods implemented in Isocompy are described in detail in section 3.2. Nevertheless, it is worth mentioning that users with coding knowledge can add other methods of their own.

Model selection steps are also implemented in two stages of the algorithm. To find the best model, the algorithm includes and combines cross validation, brute force hyperparameter searching, R-squared fitness and goodness of fit to the GMWL or local meteoric water line (LMWL), as shown in the violet boxes in Fig. 3.

Finally, the estimation step is performed in the first and second stages, as illustrated in the blue 1.4 and 2.4 boxes of Fig. 2, by determining the substantial features determined in previous steps. This workflow ensures the flexibility of the input features, time steps and geospatial scale and, at the same time, promotes and speeds up the model generation process in an integrated algorithm.

### **3.3.2 Isocompy architecture**

The Isocompy tool examines the relationship among the input variables with various linear and nonlinear regression methods, performs a statistical analysis and dimensionality reduction, and chooses the best available regression method and its respective parameters via calibration and evaluation techniques. This is done by implementing novel machine learning, data management and statistical analysis libraries such as pandas <sup>176</sup>, geopandas <sup>177</sup>, numpy <sup>178</sup>, pylr2 <sup>179</sup>, statsmodels <sup>180</sup> and scikit\_learn <sup>73</sup>. Isocompy generates extensive reports alongside figures and maps to facilitate the procedure of statistical water isotope modelling and support the user in interpreting and evaluating the results.

In this section, we describe the architecture of the underlying components and the outputs of Isocompy. It is built into six classes and 18 methods, as shown in Fig. 4. A list of the Python libraries used in Isocompy can be found in section 3.6.1.

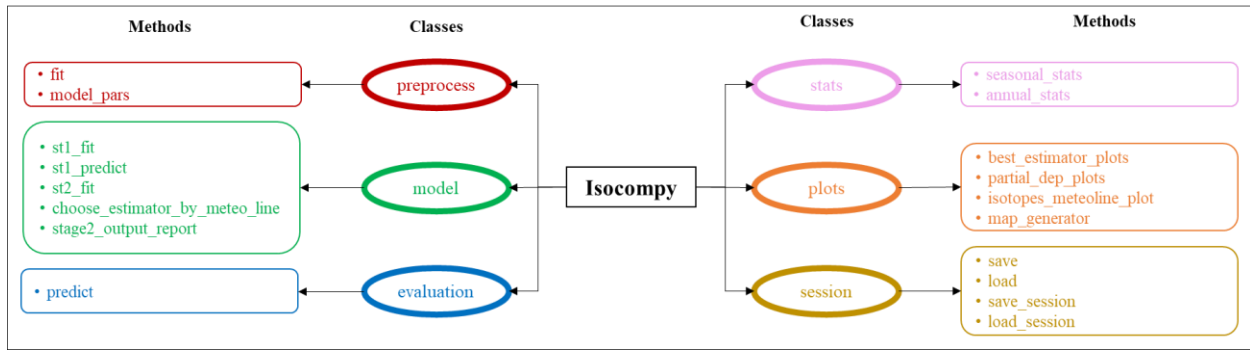


Figure 13. The Isocompy algorithm architecture. It contains 6 classes and 18 methods.

## Preprocessing

The *preprocess* class holds the data preparation step (red frames in Fig. 4), whose inputs are pandas dataframes. This class has the ability to filter outliers based on upper and lower limit percentiles or modified IQR functions<sup>181,182</sup>. Outlier detection can be performed with or without zero values included in the data removal procedure. This is possible since there are geospatial states where zero values can result in unreasonable outlier filtration (e.g., removing the 5% lowest precipitation values from an arid zone with very few precipitation events). Data averaging can be performed based on arithmetic or geometric averaging. It is also possible to define specific time periods and limit the outputs to these episodes. Another decisive feature of *preprocess* is that the user specifies the brute-force search hyperparameters of the corresponding regression models. This selection is closely dependent on the format (i.e., volume and quality of the data) and correlation of the dataset<sup>183</sup>. Therefore, it is crucial that the user have an experienced-based, focused, theoretically sound, and practical search approach. Nevertheless, the default values which are described in detail in section 3.6.1, could be useful for dealing with complex datasets in our experience.

The *model* class (green frames in Fig. 4) is designed to handle the statistical analysis, feature selection, model regression and model selection procedures in the first and second stages; the flowchart of this stage is shown in Fig. 5, and it can be performed manually or automatically. The statistical analysis and feature selection parameters are defined as arguments of the class. In the manual mode, the statistical analysis data are shown, and the user must choose the considerable features. In the automatic mode, Isocompy finds the parameters with the most influence on the dependent variable by comparing their correlation coefficients and VIFs with predefined thresholds in an iterative process. However, the usage of correlation coefficients, VIFs or threshold values can also be defined by the user. The output features of this statistical analysis and feature selection step (Fig. 5) feed the regression models.



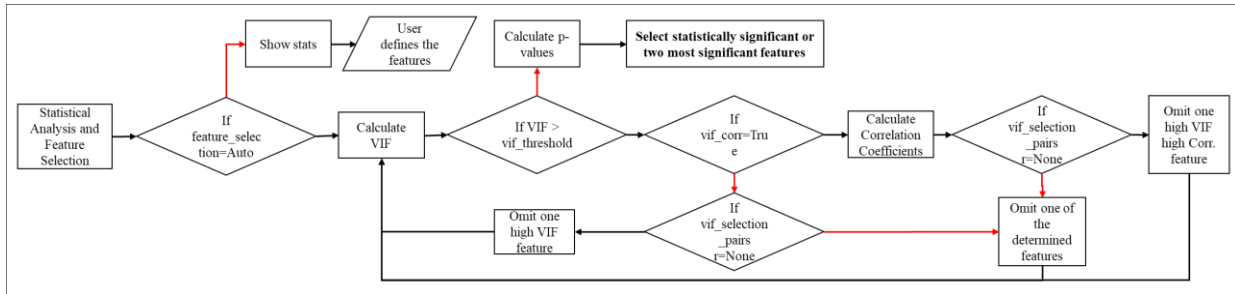


Figure 14. Feature selection flowchart of the model class. Red lines indicate false arguments.

Fig. 6 illustrates the workflow of the regression modelling, model evaluation and calibration processes that result in selecting the best model. In each regression method, all combinations of hyperparameters are defined. For each combination, the random k-fold cross-validation technique is used to avoid overfitting. The score of a determined hyperparameter set is defined as the average score of the k models. The selected set of hyperparameters for each model is defined as the one with the highest average score. The best model among different regression methods can be selected based on preferred criteria. In the first stage, the best models are selected based on higher R-squared values, whereas in the second stage, the best model can also be selected based on three different criteria: the smallest point-to-point estimation-observation distance, the pair of models with the most similar results to the LMWL or the pair of models with the most similar results to any defined line between the water isotopes. The predefined arguments for this line are eight and ten coefficient and intercept values, respectively, that represent the GMWL. To test the different options available for selecting the best model in the second stage, it is possible to change the criteria and generate corresponding outputs.

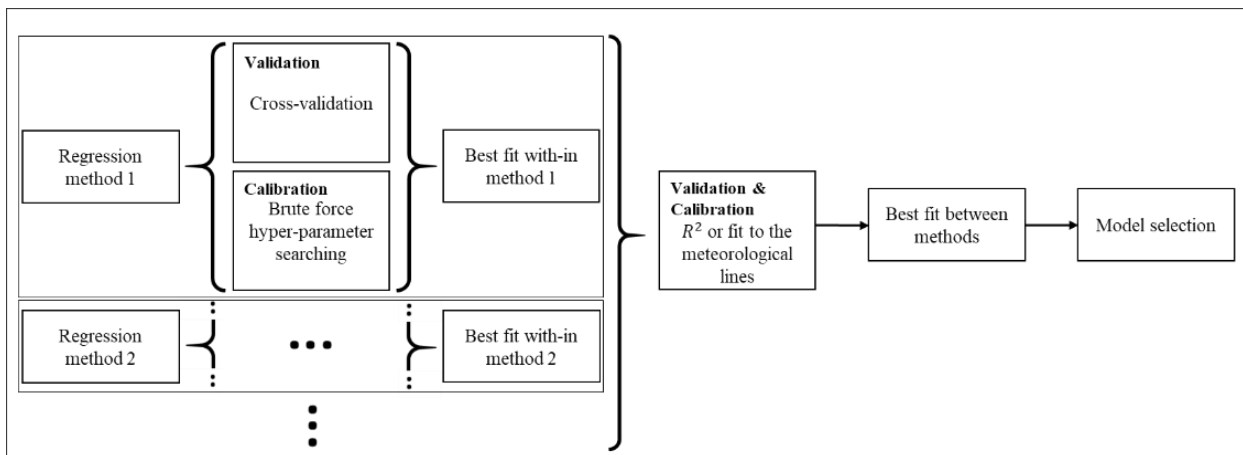


Figure 15. Workflow of the model regression, model validation, model calibration and best model selection processes. Black dots show that these processes are performed for each regression method selected.

Isocompy generates reports that include the details of all the executed models and the selected models with their R-squared values, adjusted R-squared values, VIF values, correlation coefficients, mutual information, chosen input features and sets of hyperparameters. For the chosen regression models, Isocompy also reports the cross validation averages and standard deviations obtained on the training and test data to evaluate the model estimation uncertainties.

### Model evaluation

The *evaluation* class follows the algorithm shown in Fig. 7 to calculate the outputs of the second-stage estimations. All the independent features introduced in data preparation go through the statistical analysis, and only the substantial features are used in the regression models to obtain the desired spatial-temporal estimations. Only the samples with independent determined features must be introduced, while all other data are ignored in the isotopic estimation process.

The indirect input features (box a in Fig. 2) go through the stage one estimation procedure, which are unified with the introduced direct features and are used as isotopic composition input variables for the final isotopic estimation.

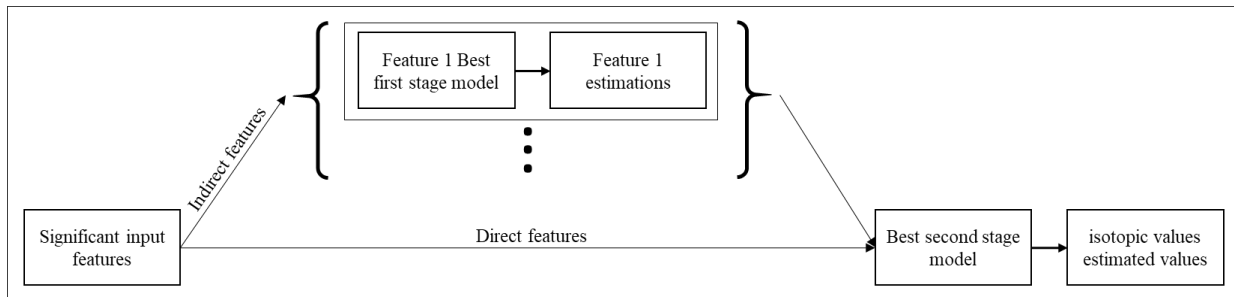


Figure 16. Workflow of the evaluation class for estimating the second-stage regressions.

### Postprocessing tools

The *stats* class generates statistical reports for each stage (pink frames in Fig. 4). They include the characteristics and details of all executed models and the selected models: their R-squared s, adjusted R-squared values, VIF values, correlation coefficients, mutual information, chosen input features and sets of hyperparameters. Isocompy also reports the chosen regression models, cross validation averages, and standard deviations of the training and test data to evaluate the model estimation uncertainties. The reports can be generated for the whole or the separate parts of the time series.

The *plot* class generates diverse kinds of graphics to illustrate the results (orange frames in Fig. 4). The *shapely*<sup>184</sup>, *Bokeh*<sup>185</sup> and *matplotlib*<sup>109</sup> libraries are employed to develop the methods of this class. The *partial\_dep\_plots* method generates partial dependency plots. The *best\_estimator\_plot* method constructs

the plots of the best estimator in each determined time window. The *isotope\_meteoline\_plot* method is designed to illustrate and compare the output data and observed data with the GMWL and LMWL. This method uses the reduced major axis (RMA) regression method to calculate the local line of the input data. It is shown that the RMA approach explains water isotope relationships better than least-squares regression since it takes the measurement errors in box axes into account<sup>186-188</sup>. The *isotope\_meteoline\_plot* method can also generate residual plots of each isotopic station for each isotopic composition and accompanies them with a report including the mean absolute errors, mean square errors and means and standard deviations of the residuals, observations and estimations.

The *map\_generator* method generates maps of the desired features, whether they are observed or estimated. The maps are generated based on the estimated data limits introduced by the user to the *evaluation class* in the time periods defined by the user. The results can be limited to positive values and/or to percentages if needed. The user has the ability to add a desired shapefile to the maps, display the measured data and define the aesthetics. The results can be saved as an interactive HTML file or in an image format.

### **Project management**

The *session* class enables the functionality of saving and loading one or all defined objects of a session (yellow frames in Fig. 4). The session class is powered by the Dill python library<sup>189</sup> because of its capacity to save the executed Isocompy project as a compressed file along with its results in a single command. Hence, it would be feasible to save and close an interpreter session, send the compressed session file to another computer, open a new interpreter, decompress the session and thus continue from the point of work saved in the original interpreter session.

### **3.3.3 Outputs**

Isocompy outputs can be categorized into four groups: reports, figures, maps and datasheets. It is possible to obtain this information at different steps to clarify the underlying processes. Reports are generated to address the input data characteristics, partial and whole time period statistics, the best first- and second-stage model characteristics, all models in the first and second stages, the best second-stage model selection scoring details based on the chosen function, prediction model uncertainty statistics, residuals, observed and estimated isotopic value statistics and errors.

Figures can be created for partial dependencies, observed-estimated regressions, residual plots and meteoric line plots, as explained in section 3.3.2.3. The bottom-left and top-right parts of Fig. 8 show examples of partial dependencies and residual plots, respectively. Examples of observed-estimated plots can be seen in the figures of the next section.

Maps can be created in different formats for any desired feature by using the *map\_generator* function, as mentioned in section 3.3.2. Examples of maps can be seen in the figures of the next section. The bottom-right part of Fig. 8 shows a screenshot of an interactive map created by Isocompy.

Datasheets are produced in the data preprocessing stage, and they include outlier-removed data, monthly averages for each year at each station and station averages. First- and second-stage estimations are also saved in datasheets. Refer to section 3.6.2 for an example datasheet.

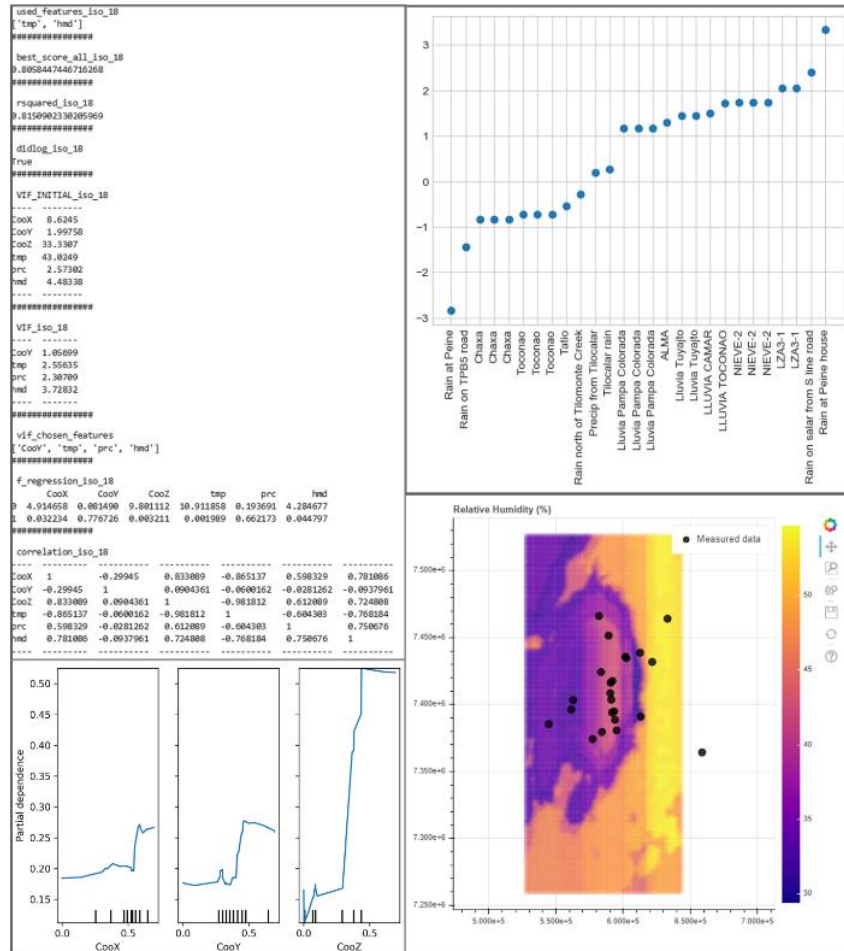


Figure 17. Screenshots of the outputs generated by Isocompy. Top left: An example report. Bottom left: Partial dependency plots of the selected features. The values are standardized between zero and one. Vertical ticks on the x-axis illustrate the percentile of the data. Top right: A residual plot at each observation point generated by Isocompy via the *isotope\_meteoline\_plot* method. Bottom right: An interactive map generated by Isocompy without an available shape file.

### 3.4 APPLICATION TO THE EXAMPLE OF SALAR DE ATACAMA

The Salar de Atacama is the ideal target zone for demonstrating Isocompy capabilities due to its particular climate and topographic features. The scope of this investigation is not a comprehensive isotopic analysis, as it has been published already <sup>190–197</sup>, but rather validate Isocompy performance. Therefore, using the

scarce information that is currently available, the climatic characteristics and isotopic composition of the precipitation in this area are compared with that of previous studies.

The Salar de Atacama basin is located in northern Chile in the Antofagasta region (Fig. 9). This zone is the largest salt flat in Chile and the third-largest salt flat in the world. The Salar de Atacama is one of the driest places on the Earth's surface, contains vast amounts of lithium reserves and is a valuable lagoon ecosystem (RAMSAR). For these reasons, in recent decades, many studies have been carried out on the water resources of this area <sup>190–197</sup>. No continuous monitoring is performed on individual precipitation events in the basin, and the available data do not have a high spatial density.

The distribution of isotopic precipitation samples is heterogeneous in time, space and type of sample. Specific rain samples are taken in the basin, and permanent rain collectors are installed close to automatic meteorological stations <sup>198</sup>. Isotopic samples are mostly collected during the summer months (January, February and March) since this is the period containing important precipitation events (during the so-called “Altiplanic winter”). As a result, Isocompy is applied only during these time periods.

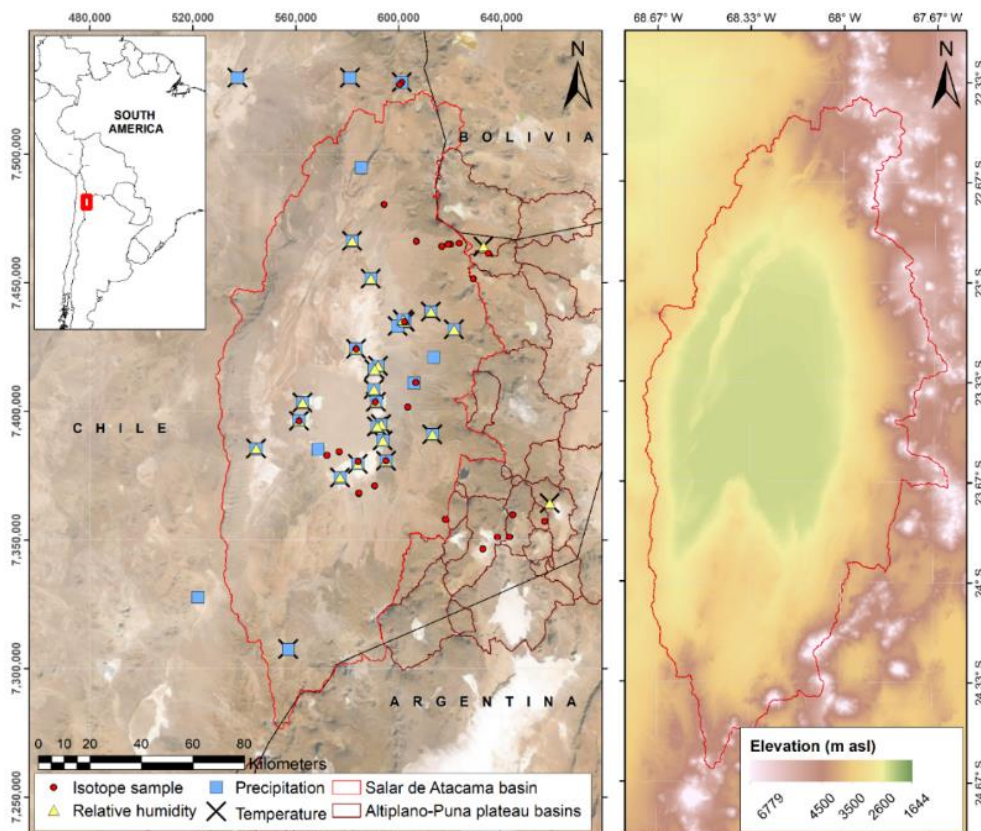


Figure 18. Left: location map of the study area in South America with published isotopic precipitation data (red circles) and automatic weather stations that monitor temperature (crosses), precipitation levels (blue squares) and relative humidity (yellow triangles). The solid red line delineates the Salar de Atacama basin, and the solid brown line shows the Altiplano-Puna plateau basins. The base map is derived from satellite data (SRTM from

<http://earthexplorer.usgs.gov/>). All location data are in UTM Zone 19 S coordinates based on the WGS of 1984. The utilized DEM is an ALOS PALSAR RTC product that has a resolution of 12.5×12.5 m and is provided by the Alaska Satellite Facility. Right: elevation map of the Salar de Atacama basin.

### 3.4.1 Input data

The available meteorological variables that potentially influence the  $\delta^{18}\text{O}$  and  $\delta^2\text{H}$  values in this study are air temperature, relative air humidity and amount of precipitation. They are recorded daily at the meteorological stations of the Salar de Atacama basin and its surroundings and compiled from the automatic weather stations belonging to the General Directorate of Waters <sup>199,200</sup> of Chile and the Soquimich (SQM) mining company. Temperature value records are provided from 28 stations for the period from 1974 to 2019, ranging from -5.5°C to 22.6°C (mean 16.3°C), relative humidity values are derived from 24 stations from 1987 to 2019, ranging from 2.3% to 76.3% (mean 33.3%) and precipitation volume data come from 31 meteorological stations from 1959 to 2019, ranging from 0mm to 219mm (mean 14.9mm).

The 52 precipitation samples for the isotopic analysis are compiled from previously published studies <sup>201-206</sup> from 2002 to 2021, ranging from -18.9 ‰ VSMOW to 2.5 ‰ VSMOW (mean -7.0 ‰ VSMOW) and from -139.7 ‰ VSMOW to 21.7 ‰ VSMOW (mean -2148.3 ‰ VSMOW) for  $\delta^{18}\text{O}$  and  $\delta^2\text{H}$  respectively and correspond to 31 different points (Fig. 9). These samples are heterogeneous: some are individual precipitation events, others are monthly accumulated or multimonth samples, and others are mixtures of rainfall and snow. Refer to section 3.6.2 for the input data file.

### 3.4.2 Implementation

The data preparation steps for the input parameters are shown in Fig. 10 Lines 7 to 25 align with the *preprocess* classes for precipitation, air temperature and cumulative humidity. These three indirect variables are estimated in stage one and are dependent on the spatial variables (latitude, longitude and altitude).

Lines 26 to 32 create the *preprocess* class for the  $\delta^{18}\text{O}$  and  $\delta^2\text{H}$  values of precipitation used in the second stage of the model. The spatial variables here act as direct variables since they are measured in the same location as the isotopic data. In this study, outlier removal techniques, such as those explained in section 3.3.1, are not needed.

```

1. #Import isocompy
2. from isocompy.data_preparation import preprocess
3. from isocompy.reg_model import model
4. from isocompy.tools import stats, plot
5. #-----
6.
7. #Data preparation: rain, temp and hum are pandas DataFrames, imported from the database.
8. dir = "defined directory"
9. fields=["CooX", "CooY", "CooZ"]
10. #-----
11.
12. #Precipitation preprocess class
13. pre_prc=preprocess()
14. pre_prc.fit(inp_var=rain,var_name="prc",fields=fields,remove_outliers=False,direc=dir)
15. #-----
16.
17. #Temperature preprocess class
18. pre_tmp=preprocess()
19. pre_tmp.fit(inp_var=temp,var_name="tmp",fields=fields,remove_outliers=False,direc=dir)
20. #-----
21.
22. #Humidity preprocess class
23. pre_hmd=preprocess()
24. pre_hmd.fit(inp_var=hum,var_name="hmd",fields=fields,remove_outliers=False,direc=dir)
25. #-----
26.
27. #isotopes
28. pre_iso1=preprocess()
29. pre_iso1.fit(inp_var=iso_18,var_name="iso_18",fields=fields,remove_outliers=False,direc=dir)
30.
31. pre_iso2=preprocess()
32. prep_iso2.fit(inp_var=iso_2h,var_name="iso_2h",fields=fields,remove_outliers=False,direc=dir)

```

Figure 19. Isocompy data preparation. Location information (X, Y: coordinates; Z: altitude) is used to calculate the feature information in these positions. *Preprocess* classes are created for the precipitation, temperature, cumulative humidity,  $\delta^{18}\text{O}$  and  $\delta^2\text{H}$  of precipitation. *Rain*, *temp* and *hum* are panda dataframes that contain *ID*, *Date* and *Value* columns.

The steps needed to execute the first-stage models for the desired *preprocess* classes in January, February and March of all years can be seen in Fig. 11. Each *preprocessing* class contains the dependent and independent variables, as shown in Fig. 10. The models for each process class and for each month are isolated from the rest. The feature selection options of the first stage are not changed from the predefined default Isocompy values (line 5 in Fig. 11), so the feature selection process in the first stage runs automatically. Isocompy reports the VIF and correlation coefficient values but does not consider them in the feature selection procedure. Executing lines 9 and 10 generates the estimated-versus-observed values and partial dependency plots for each regression model in stage one.

```

1. #stage 1 model class
2. dir ="defined directory"
3.
4. est_class=model()
5. est_class.st1_fit(var_cls_list=[pre_prc,pre_tmp,pre_hmd],st1_model_month_list=[1,2,3],direc=dir)
6. #-----
7.
8. #stage 1 model plots
9. plots.best_estimator_plots(est_class,st2=False)
10. plots.partial_dep_plots(est_class,st2=False)

```

Figure 20. Stage-one estimation models, estimator and partial dependency plots.

To create the second-stage models, the precipitation, temperature and humidity values must be predicted at the same coordinates as the isotopic measurements. Line 2 in Fig.12 3 estimates these values for three months based on the stage-one models. The dependent and independent (direct and indirect) variables must be determined as shown in lines 7 to 10.

The feature selection process of the second-stage models is the *st2\_fit* method, which is performed automatically if it is not specified. In this example, some aspects of the feature selection process are specified. Thus, as seen in line 13 of Fig. 12, in cases with high VIF and correlation coefficient values, one of the parameters is removed: temperature is preferred over altitude to respect the seasonality of the data. Line 16 executes the model based on the defined variables, and lines 19 to 23 generate the statistical reports of the given month and the whole period. Similar to the first stage, lines 26 and 27 generate the estimated-versus-observed values and partial dependency plots for each generated isotopic regression model.



```

1. #Stage 1 models prediction
2. est_class.st1_predict(cls_list=[pre_iso1,pre_iso2],st2_model_month_list=[1,2,3])
3. #-----
4.
5. #Stage 2 model
6.
7. #Determine the dependent and independent variables - direct ("CooX","CooY","CooZ") or indirect
   ("tmp","prc","hmd") - to take into account for each model in the second stage
8. st2_model_var_dict={
9.     "iso_18":["CooX","CooY","CooZ","tmp","prc","hmd"],
10.    "iso_2h":["CooX","CooY","CooZ","tmp","prc","hmd"]}
11.
12. #Defining that taking into account vif and correlation coefficients, if the algorithm has to remove
   one of the variables between the "CooZ","tmp" pair, it has to be "CooZ"
13. args_dic={"vif_selection_pairs":[["CooZ","tmp"]]}
14.
15. #Stage 2 model fit
16. est_class.st2_fit(model_var_dict=st2_model_var_dict,args_dic=args_dic)
17. #-----
18.
19. #monthly statistics
20. stats.monthly_stats(est_class)
21.
22. #whole period statistics
23. stats.seasonal_stats(est_class)
24.
25. #Stage 2 model plots
26. plots.best_estimator_plots(est_class,st1=False)
27. plots.partial_dep_plots(est_class,st1=False)

```

Figure 21. Stage-one estimation calculations (line 2). Stage-two model argument definitions (lines 7-10). Stage-two model execution (line 16). Statistical reports and plots (lines 26-27).

The reader is referred to section 3.6.2 for the complete version of the Jupyter notebook in this study that contains the *evaluation* class, visualization options, evaluations, estimated value datasheets, meteoric lines for observed and newly defined coordinates, residual plots and feature maps.

### 3.4.3 Results and discussion

The first stage of statistical analysis shows that altitude and longitude are significant variables for temperature and relative humidity in all three months, while latitude is also significant in March (Table 1). This is consistent with the DICTUC <sup>207</sup> results. For precipitation, latitude and altitude are significant variables in the three summer months, as Houston and Harley <sup>208</sup> mentioned, while longitude is also significant in February and March. The influence of altitude on the amount of precipitation that falls in the eastern part of the basin is recognized by all existing studies <sup>193,209,210</sup>.

Monthly models for temperature, relative humidity and precipitation are created by using the significant features. The estimation method with the highest scores in all models is the random forest, whose R-squared values are shown in Table 1. Column  $Ln(x+1)$  shows the models whose feature  $Ln(x+1)$  values are used since they result in higher R-squared values.

Table 4 Results of the first-stage statistical analysis and models per month. The bold p values denote significant parameters (<0.05).

Month	Dependent feature	ρ-value			R <sup>2</sup>	Standardized Standard deviation*	Ln (x+1)
		Longitude	Latitude	Altitude			
January	Temperature	<b>7.04E-03</b>	6.96E-02	<b>8.06E-18</b>	0.98	0.23	No
	Relative humidity	<b>1.10E-05</b>	7.06E-02	<b>1.10E-05</b>	0.87	1	Yes
	Precipitation	2.34E-01	1.73E-01	<b>7.15E-11</b>	0.97	0.09	Yes
February	Temperature	<b>1.68E-03</b>	1.16E-01	<b>1.18E-20</b>	0.98	0.28	No
	Relative humidity	<b>8.09E-03</b>	9.58E-02	<b>7.01E-03</b>	0.84	0.13	No
	Precipitation	<b>6.67E-03</b>	<b>1.52E-02</b>	<b>5.74E-08</b>	0.96	0.68	Yes
March	Temperature	<b>1.80E-02</b>	<b>2.87E-02</b>	<b>2.05E-17</b>	0.99	0	No
	Relative humidity	<b>6.38E-04</b>	<b>1.14E-02</b>	<b>3.32E-04</b>	0.82	0.78	No
	Precipitation	2.55E-01	<b>1.03E-03</b>	<b>2.00E-06</b>	0.94	0.09	No

\* Standardized standard deviation of the cross-validation scores of the estimation models.

The estimation uncertainties can be evaluated by the standardized standard deviation of the cross-validation scores for the randomly selected test dataset in each iteration (Table 1). The limited spatial distribution of the available data in the Salar de Atacama basin can play an important role in high estimated standard deviation values obtained for some features. Fig. 13 shows the observed-versus-estimated values of the three features in three months.

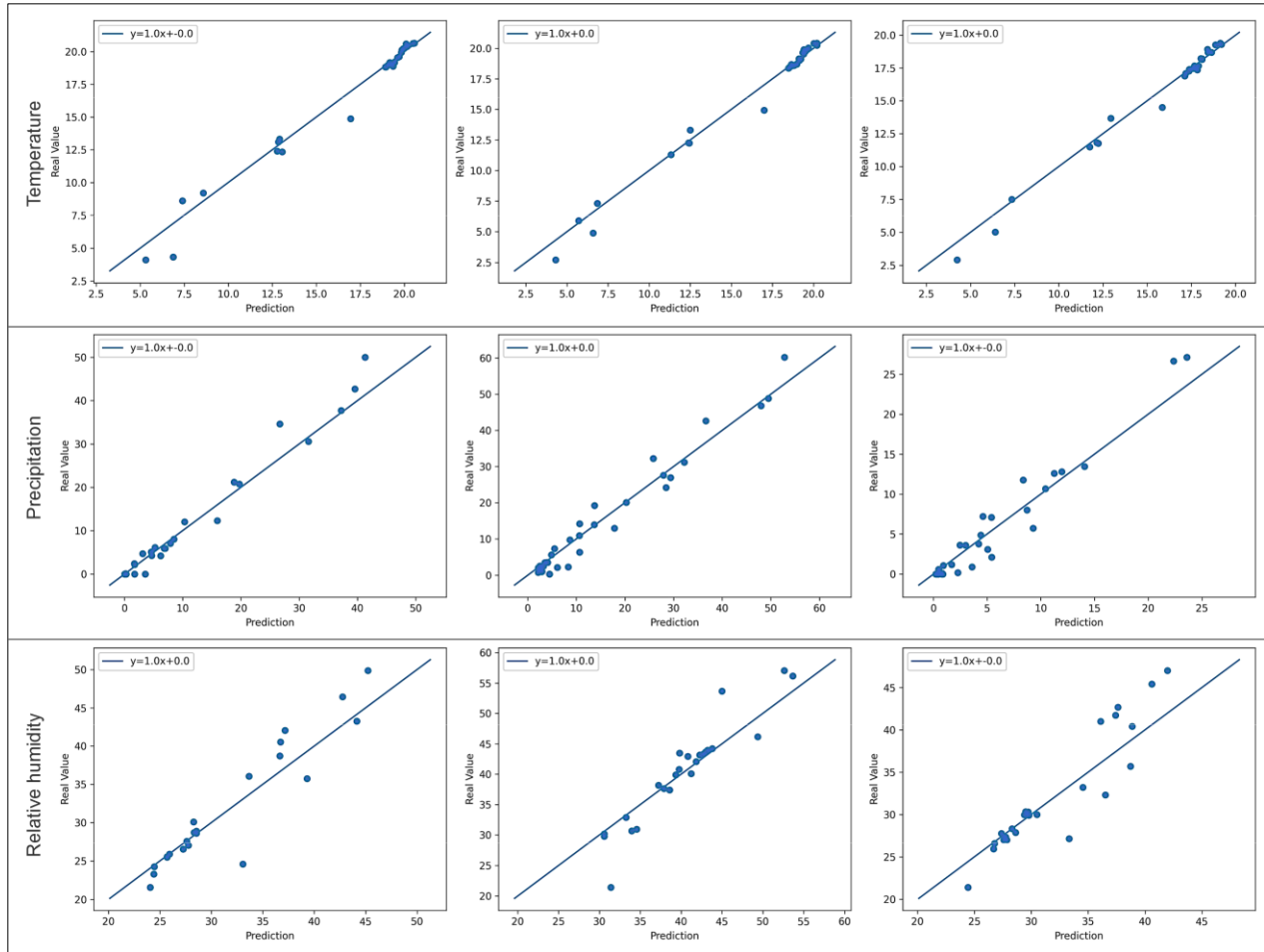


Figure 22. Plots of the estimated-versus-observed values generated by Isocompy for temperature, precipitation and relative humidity in January, February and March.

The map of the temperature distribution estimated by Isocompy for the Salar de Atacama in the three summer months is shown in Fig. 14. It can be observed that the maximum temperatures are recorded in the central area with values between 19 and 20.4°C, which are slightly lower than those of Marazuela et al. (24°C) and Kampf et al. (23°C) <sup>209,211</sup> in February. It is observed that temperature decreases with altitude, reaching minimum values of 4 to 5.3°C in the volcanic arc that surrounds the eastern side of the basin, with a gradient of approximately -0.55°C/100 m. These gradients are similar to those presented by DICTUC and MOP-DGA (-0.56°C/100 m and -0.65°C/100 m, respectively) <sup>207,212</sup>.

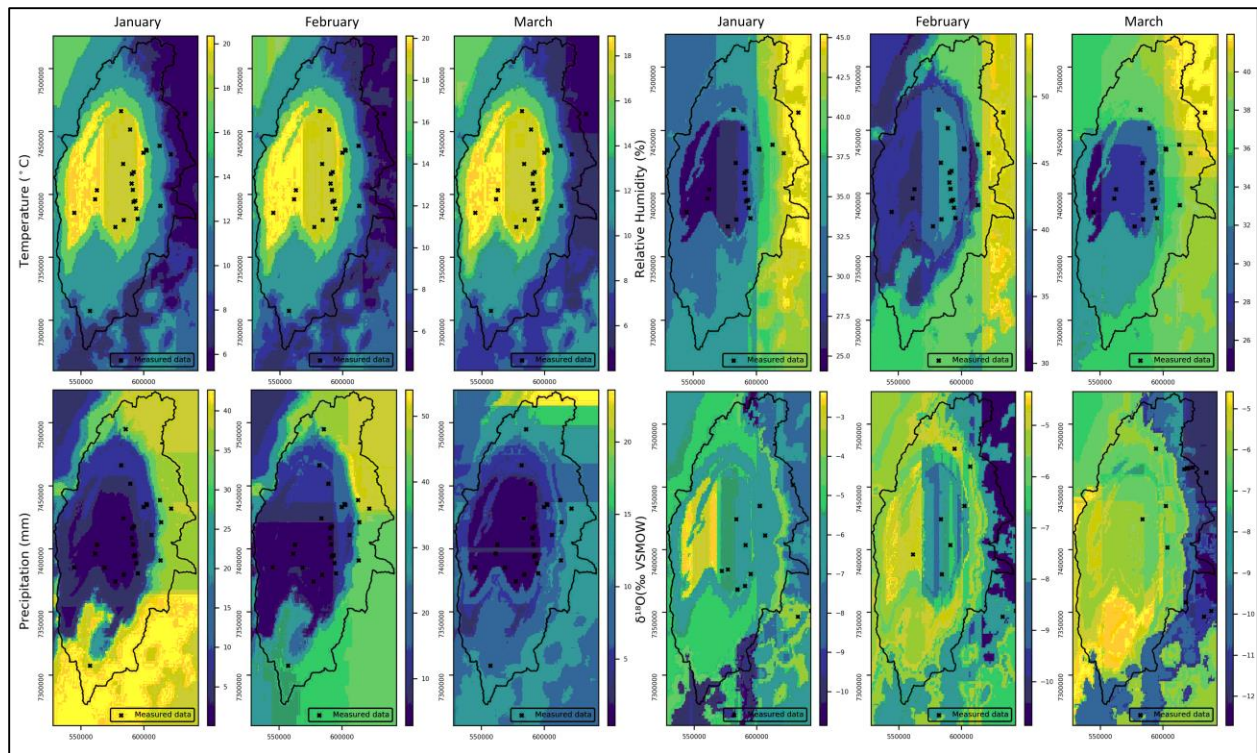


Figure 23. Maps of the temperature, precipitation, relative humidity and  $\delta^{18}\text{O}$  values of precipitation estimated by Isocompy in January, February and March in the Salar de Atacama basin.

The relative humidity values estimated by Isocompy in the Salar de Atacama basin for the three summer months can also be seen in Fig. 14. The lowest values of relative humidity are recorded in the core (24-29%) and in the west, and they increase with altitude, reaching their maximum values in the east of the basin (42-55%) and resulting in a gradient of 0.49%/100 m. In Valdivielso et al. <sup>206</sup>, who used a larger study area (N Chile), the estimated values of relative humidity in the salt flat nucleus were similar to those in the present study, although the estimated values for high altitudes were lower.

Summer storms in the Salar de Atacama basin are convective and are characterized by highly variable intensity <sup>191,193,213,214</sup>, with years that are much wetter than others and some with practically no precipitation. This high variability, accompanied by the nature of the available precipitation data, results in a low correlation between the precipitation values recorded in different seasons, as well as between the precipitation values recorded in the same season for different periods. Therefore, the precipitation models exhibit high sensitivity to anomalous values since they greatly affect the average precipitation at a station.

From the precipitation model, zero precipitation (0 mm) is estimated in the salt flat nucleus (Fig. 14), increasing with altitude up to 55 mm in the summits at the eastern limit of the basin; there is little precipitation at the western limits. A comparison with many studies that have presented annual isohyets maps of the Salar de Atacama <sup>197,207,212,215</sup> shows that the magnitude of precipitation is lower in the present

study since only the summer precipitation is considered, but the overall distribution of precipitation is similar. The summer precipitation gradient from the salt flat nucleus to the eastern peaks is 3.7 mm/100 m, which is slightly less than the annual gradients calculated in Salas et al., Valdivielso et al. (5 mm/100 m) and IDAEA-CSIC (4.6 mm/100 m)<sup>197,216,217</sup>, as these studies considered all precipitation events during the year. In contrast, DGA<sup>191</sup> calculated values of 2.7 mm/100 m in January, 2.2 mm/100 m in February and 1.8 mm/100 m in March for the period from 1970 to 2008.

Precipitation is depleted in heavy isotopes with elevation, with an average gradient of -0.19‰/100 m in summer (Fig. 14). This gradient is slightly lower than the others calculated in this region (-0.34‰/100 m in Herrera et al. and -0.26‰/100 m in Villablanca<sup>203,219</sup>). The distribution map of the stable isotopic signature is consistent with the distributions of the highest temperatures, the lowest relative humidity values and precipitation in the salt flat nucleus; at higher elevations, the precipitation and relative humidity are higher, and the temperatures are lower<sup>205,220</sup>.

In the statistical analysis and feature selection processes of the second stage, the initial VIF values are higher than the defined threshold (VIF =5) for longitude, altitude and temperature. Furthermore, these variables have high correlations with each other (Table 2). Therefore, as these variables have high multicollinearity and strong correlations, altitude and longitude are iteratively removed as important features until VIF values below the threshold are reached for all the features, as seen in the VIF\_fin column of Table 2. Then, the p values of latitude, temperature, precipitation and humidity are evaluated, and as a result, temperature and relative humidity are selected as significant features for the  $\delta^{18}\text{O}$  and  $\delta^2\text{H}$  regression models (Table 2). The R-squared values of the  $\delta^{18}\text{O}$  and  $\delta^2\text{H}$  estimation models are 0.82 and 0.79, and the standard deviations of the associated cross-validation scores are 0.58 and 0.46, respectively. The top-left and top-right plots in Fig. 15 show the estimation-versus-real measurements of  $\delta^{18}\text{O}$  and  $\delta^2\text{H}$ , respectively.

Table 5 The VIF values and correlation coefficients of the second-stage input features. VIF\_init and VIF\_fin show the initial and final VIF values, respectively. Cor. shows the correlation coefficients of the features. The p values of the parameters selected by the VIF process are shown. Significant p values are displayed in bold fonts (<0.05).

$\rho$ - valu e	VIF_fi n	VIF_ini t		Cor. Lon.	Cor. Lat.	Cor. Alt.	Cor. Temp.	Cor. Prec.	Cor. Hum.
-	-	<b>8.6</b>	Lon.	1.00	-	-	-	-	-
<b>0.78</b>	<b>1.1</b>	2.0	Lat.	-0.30	1.00	-	-	-	-
-	-	<b>33.3</b>	Alt.	<b>0.83</b>	0.09	1.00	-	-	-
<b>0.00</b>	<b>2.6</b>	<b>43.0</b>	Temp.	<b>-0.87</b>	-0.06	<b>-0.98</b>	1.00	-	-
<b>0.66</b>	<b>2.3</b>	2.6	Prec.	0.60	-0.05	0.61	-0.60	1.00	-
<b>0.04</b>	<b>3.7</b>	4.5	Hum.	0.78	-0.09	0.72	-0.77	0.75	1.00

The LMWL is calculated with the isotopic measurements (observed LMWL: yellow line in Fig. 15), the average estimated  $\delta^{18}\text{O}$  and  $\delta^2\text{H}$  values at the same points as the measurements (estimated LMWL in Fig. 15; bottom left) and the average estimated  $\delta^{18}\text{O}$  and  $\delta^2\text{H}$  values in all the study areas (estimated LMWL in Fig. 15; – bottom right). Based on the estimated LMWL at the observation points, the isotopic model has slightly different slope (7.5) and intercept (7.8) values than those obtained with the LMWL defined in different areas of northern Chile <sup>218,221–225</sup>. However, these differences are expected since the LMWL is calculated based on a different group of points in a larger area. Fig. 15 also demonstrates that the slope and intercept of the estimated and observed LMWLs are similar, which indicates that the estimated isotopic values have the same behaviour as the measured values that validates the statistical built-in capabilities of Isocompy.

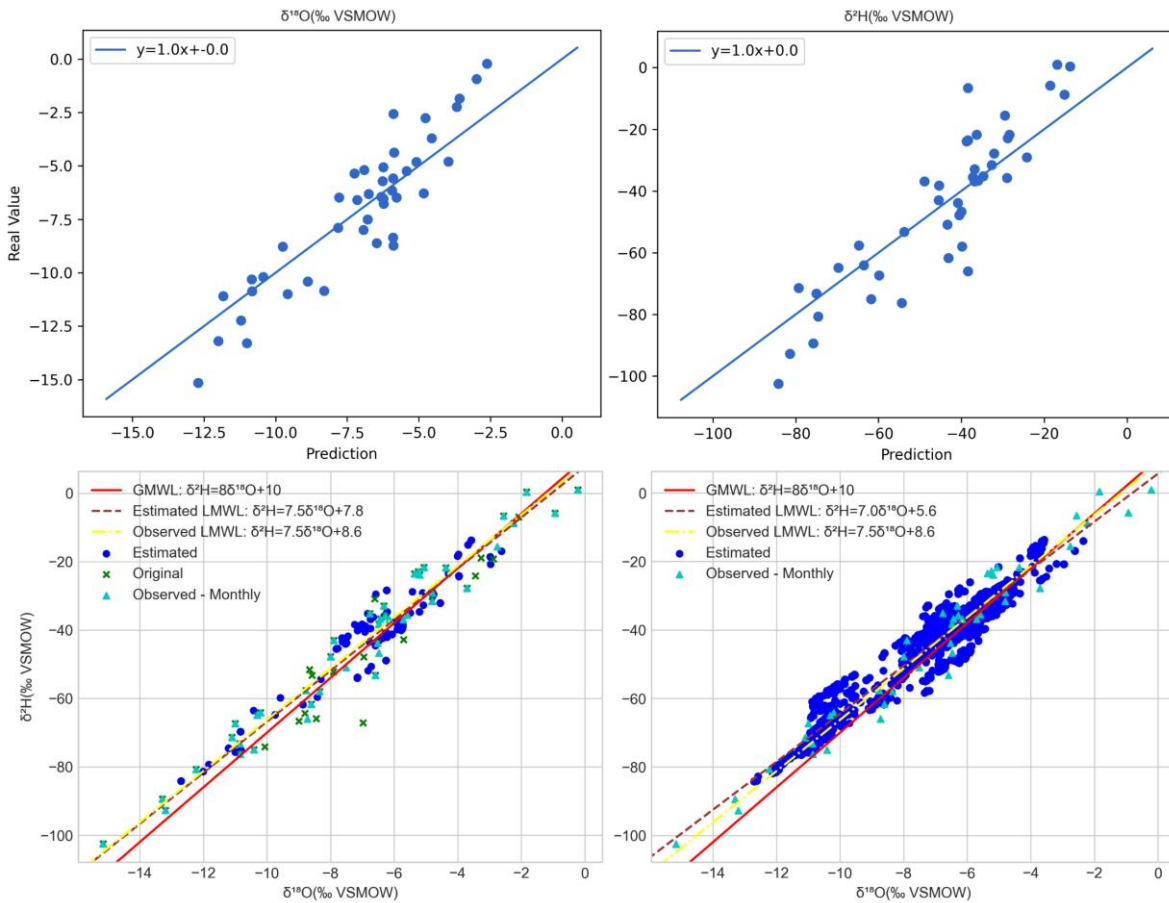


Figure 24. Top left and top right: estimations versus the measurements of  $\delta^{18}\text{O}$  and  $\delta^2\text{H}$ , respectively. Bottom: plots of the estimated (circles) and observed (triangles)  $\delta^{18}\text{O}$  versus  $\delta^2\text{H}$  values of precipitation. The red line is the GMWL, the brown dashed line is the estimated LMWL, and the yellow dashed line is the observed LMWL. Bottom left: the plot obtained using the  $\delta^{18}\text{O}$  and  $\delta^2\text{H}$  values estimated at the same points as the measurements. Bottom right: the plot obtained using the  $\delta^{18}\text{O}$  and  $\delta^2\text{H}$  values estimated in the study area. The reader is referred to section 3.6.2 for the monthly meteoric line plots, residual plots and reports. All plots are generated by Isocompy.

### 3.5 CONCLUSION

Isocompy is an open-source Python library dedicated to regression, statistical analysis and modelling for isotopic compositions of natural water. It considers the features that potentially affect the isotopic signature in a multistage procedure. These features can be meteorological measurements, particle trajectory-related parameters, sea surface temperatures, variables derived from reanalysis or any other parameter desired by the user.

The code simplifies and optimizes the analyses of the isotopic characteristics of natural water. The isotopic composition obtained using the Isocompy applications are consistent with those obtained in previous studies in the Salar de Atacama, which was used as an example of a study area with scarce and heterogeneous data, for validation. Therefore, Isocompy is capable of producing accurate estimated isotopic spatial distribution and estimated LMWL data. The application of Isocompy in this complex area (with unequal datasets in space and time) demonstrate the versatility on using machine learning techniques in environmental studies. Isocompy can deliver reasonable outputs, accompanied by an automatic feature selection procedure that enables a fast yet extensive study of the features that affect the isotopic composition of precipitation. The easily generated statistical analysis reports, feature maps and meteoric line plots from the observed and estimated values make the evaluation process simple and user friendly.

Nevertheless, choosing the right set of regression methods and defining a suitable set of hyperparameters for each method in a specific study area, considering the available computation power and time, is always challenging, as is selecting a suitable time window. In cases with high data densities, the number of regression models in the first stage of Isocompy can be increased by shortening the time window of each model and proceeding with the same time window in the second stage. In contrast, similar to the example of the Salar de Atacama, when the data do not have high density, it is possible to widen the time window and use data integration techniques to include more input data in the first stage, integrating the different first-stage outputs into a single model in the second stage.

Another important aspect to consider is the sensitivity of the models to anomalies in the input data. This effect is more visible when the data are scarce. Data treatment techniques such as outlier detection and data filling can be effective with Isocompy in decreasing the sensitivity of the models, but the anomalies must be considered when interpreting the results.

Although Isocompy focuses mainly on the isotopic composition of precipitation, the code could assist researchers to further environmental investigations such as paleoclimate change studies which obtaining the environmental variables from stable isotopes could be challenging. In studies where data preprocessing, statistical analysis, feature selection and machine learning are needed to investigate an environmental

feature, Isocompy can be an integral solution for facilitating the workflow. In addition, Isocompy is an open-source library in a widely used programming language, which makes it a good candidate for further additions/implementations and customizations in different study areas.

Isocompy is a flexible tool that can be adapted based on the amount of data available in time and space, and it has the capability to apply diverse regression methods. It provides the user with reports, figures, datasheets and maps to facilitate the comprehension of the underlying process of each step and to speed up isotopic composition studies. Isocompy is designed to be easy to use but at the same time maintain adaptability to different studies.

### **3.6 SOFTWARE AND DATA AVAILABILITY**

The datasets generated and analysed during the current study are available in the GitHub repository, <https://github.com/IDAEA-EVS/Isocompy> under AGPL-3.0 license.

#### **3.6.1 Isocompy library information**

Year first available: 2022

Dependencies: pandas, pylr2, dill, geopandas, bokeh, statsmodels, numpy, tabulate, matplotlib, Shapely, scikit\_learn

Contact information: [ashkan.hassanzadeh@csic.es](mailto:ashkan.hassanzadeh@csic.es).

Refer to <https://github.com/IDAEA-EVS/Isocompy/wiki> for additional information about the installation, default values of the arguments, explanation and the usage.

#### **3.6.2 Application on Salar de Atacama**

The input data, the output reports, plots, figures and maps alongside the Jupyter notebook are available free of charge in <https://github.com/IDAEA-EVS/Isocompy>.



## 4 An open source Python library for water balance modelling

This chapter is based on:

**Hassanzadeh, A., Vázquez-Suñé, E., Valdivielso, S., Corbella, M.** An open source Python library for water balance modelling. *Environ. Environmental Modelling & Software* (2023). (*submitted*)

## 4.1 INTRODUCTION

As the global population continues to grow and climate change intensifies, sustainable groundwater management becomes increasingly crucial for ensuring long-term access to safe and reliable water resources.<sup>226</sup> One of the key aspects of groundwater management is groundwater recharge assessments, which quantify precipitation and other water resources that enter groundwater reservoirs (aquifers). However, many methods have been used to estimate recharge, choosing the appropriate method is of great importance<sup>21</sup>. A common approach for recharge estimation is based on the source of information employed, such as surface water, unsaturated and saturated zone techniques, empirical formulas, or a mix of these methods<sup>22</sup>.

Surface water-based approaches generally focus on the relationship between the aquifer and surface water dynamics, which is determined by soil characteristics<sup>227</sup>. Unsaturated zone-based techniques estimate groundwater recharge based on the drainage below the root zone<sup>15,23,228,229</sup>. Recharge estimations using saturated zone-based techniques are generally derived from observed data in saturated zones such as the groundwater level measurements<sup>230</sup>.

Advancements in computation power and ability have resulted in the design of many computer programs that can aid experts in recharge assessment. The programs for recharge assessment are based on different methods and useful in different manners, but these programs have some shortcomings that have not been fully addressed to date:

1. Homogeneity of spatial data. Depending on the objectives of the study and data availability, groundwater recharge assessments vary greatly in scale, from regional to local studies<sup>15</sup>. Some programs have limited spatial variability (Easy Bal)<sup>24</sup>.
2. The time window. Different time windows (hourly, daily, monthly, etc.) have to be used depending on the objectives of the study, data availability, and so forth. Computer programs bound by a specific time window can be potentially limited in this aspect (SWAT)<sup>25</sup>, (HYDROBAL)<sup>231</sup>, (SWB)<sup>232</sup>, (HydroBudget)<sup>27</sup>.
3. Input data flexibility. If data of a study area (measured or calculated) are already available, users should be able to introduce them to the program and use them during the assessment to prevent recalculation of parameters (VISUAL-BALAN V2.0 and GIS-VISUAL-BALAN)<sup>28</sup>.
4. The complexity of urban infrastructure. In regional studies, it is often crucial to consider the urban water cycle, including the complexities of the urban infrastructure, which not all computer programs are capable of (WetSpass)<sup>29</sup>, (WRF-Hydro)<sup>30</sup>.
5. Water recharge calculation methods. Water recharge-related parameters can be calculated using various methods. Some programs offer very limited options to the user at this stage (VISUAL-BALAN)<sup>28</sup>, (PRO-GRADE)<sup>31</sup>.
6. Open sources. Some computer programs restrict their use to subscribers (WetSpass)<sup>29</sup>, (PRO-GRADE)<sup>31</sup>, (HYDRUS)<sup>32</sup>, (HEC-HMS)<sup>233</sup>.

7. Knowledge of scripting language. A prerequisite for using some computer programs is that the user needs to be familiar with scripting languages, which considerably limits the usage of the programs to some users (WRF–Hydro)<sup>30</sup>, (SWB)<sup>232</sup>.
8. Output. To understand and evaluate the results or outputs of the study, various maps, figures, and reports are often needed. Some computer programs generate limited output files, which is time-consuming.
9. Database. Generally, a wide range of data is available in the study zone, which can be considered during water recharge assessments. Not using a well-known database with the ability to integrate these parameters can make it challenging to have a broader vision of the study (CRHM)<sup>33</sup>, (VISUAL-BALAN)<sup>28</sup>.
10. Future development. Recent advancements in computational power and machine learning algorithms suggest that computer programs have the ability to interface with modern algorithms. Some computer programs are limited in this aspect (HYDRUS)<sup>32</sup>, (HEC–HMS)<sup>233</sup>.

Therefore, this study presents WaterpyBal, a code that addresses some of the abovementioned shortcomings. WaterpyBal is an open-source modular Python library<sup>86</sup> that helps the user at different stages of the soil water balance (SWB) assessment. The program takes into account the vertical water movement and diffused precipitation and recharge. It incorporates the principles of hydrological/watershed modeling methods, and offers the following: (1) flexibility in input data, time interval, and spatial–temporal properties; (2) a collection of tools to facilitate the different stages of the study; (3) a well-known database, as the core dataset, with the ability to integrate a broad range of information that is supported by a wide series of programs; and (4) a base for future developments and contributes to reproducible research owing to its open-source and having modular design. Moreover, it is accompanied by the WaterpyBal Studio, a graphic user interface of WaterpyBal. The WaterpyBal Studio allows the use of the WaterpyBal library even if the user does not have scripting knowledge by incorporating the most common capabilities of WaterpyBal in the graphic user interface.

The following sections explain the fundamentals used in this study, followed by a description of the WaterpyBal code and workflow. To demonstrate the functionality of WaterpyBal, water recharge was calculated using WaterpyBal in a synthetic study area.

## 4.2 METHODS

There are several steps to calculate the SWB, as shown in Fig. 1. The infiltration and runoff can be calculated using the area characteristic parameters, rainfall, and irrigation (dark green boxes in Fig. 1). The potential evapotranspiration (PET) can be calculated using the parameters needed in the PET method chosen by the expert (orange box in Fig. 1). If the study area includes an urban zone, the SWB parameters have to be modified using urban area characteristics (yellow box in Fig. 1). The soil water reserve (SWR) will be calculated using the soil characteristics of the study area (light green box in Fig. 1). The SWB parameters such as recharge, real evapotranspiration (RET), deficit, and runoff will be calculated using the

parameters calculated in previous steps (light blue box in Fig. 1). The following sections describe each step of the SWB calculation in detail.

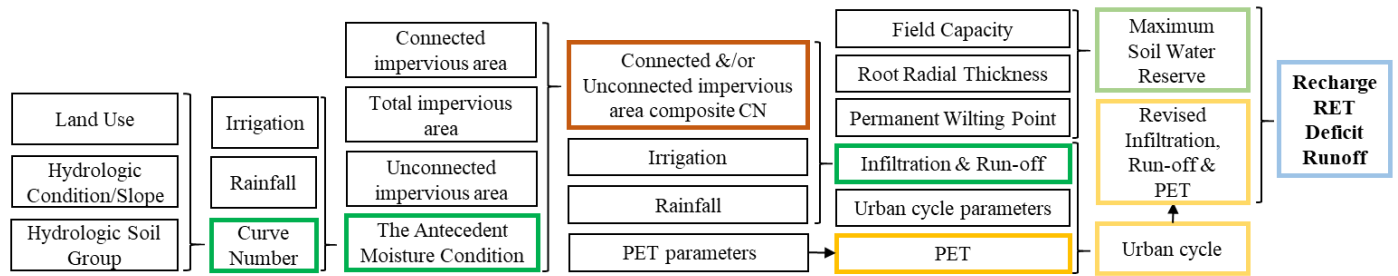


Figure 25 The general scheme to calculate the SWB. PET: Potential Evapotranspiration, RET: Real Evapotranspiration, CN: Curve Number. Each color defines a parameter that is calculated using the input data or other calculated parameters.

#### 4.2.1 Soil Water Balance

The SWB equation (Eq. 1) (known as the water budget equation) is used for recharge calculation. Considering the vertical water movement, the total rainfall in a region is the addition of the water runoff, water infiltration, and water lost via evaporation or other phenomena:

$$P = Ia + F + Q$$

Eq. 1

where P is the total rainfall; Ia is the initial abstraction that accounts for all losses before runoff occurs and generally comprises interception, surface evaporation, and surface depression; F is the cumulative infiltration to the subsurface excluding the retention from Ia; and Q is the runoff.

#### 4.2.2 Infiltration and Runoff

The Natural Resources Conservation Service curve number (NRCS-CN; shortened to CN in this study) rainfall-runoff empirical method is one of a widely used models to calculate runoff and infiltration (shown in green in Fig. 1) <sup>234-236</sup>. This method was originally designed for small agricultural watersheds in the US, but it is used in many projects around the world. Numerous studies have suggested different modifications to the empirical formula or defined the locally viable parameters for CN method formula <sup>237</sup>.

Two empirical equations (Eqs. 2 and 3) have been considered to solve Eq. 1 using the CN method.

$$\frac{F}{S} = \frac{Q}{(P - Ia)}$$

Eq. 2

$$Ia = \lambda S$$

Eq. 3

where  $S$  is the potential maximum retention and  $\lambda$  is the potential abstraction ratio that ranges from 0 to 1. Eq. 2 assumes that the ratio of the cumulative infiltration to the potential maximum retention is equal to the ratio of the runoff and water contacting the ground.

Eqs. 1 and 2 can be combined to get Eq. 4.

$$Q = \begin{cases} \frac{(P - Ia)^2}{(P - Ia + S)}, & P > Ia \\ 0, & P < Ia \end{cases}$$

Eq. 4

Eq. 4 specifies that if the precipitation is less than the calculated initial abstraction, the runoff will be 0. Several parameters have to be considered to calculate infiltration and runoff: The CN value must be determined using the existing tables that characterize the study area based on their usage, soil type, and hydrologic conditions (HCs); then, considering the rainfall and irrigation cumulative values, the antecedent moisture condition (AMC) has to be calculated and the CN values have to be corrected, respectively. In urban zones, depending on the study, the composite CN (CCN) can be calculated for connected impervious areas (CIAs) and unconnected impervious areas (UIAs). The following sections discuss the details of the aforementioned calculations using the CN method.

### Curve Number

The CN method is designed to calculate  $S$  (potential maximum retention), which can then be used to calculate  $Q$  and  $F$  using Eqs. 2 and 4, respectively. In the CN method, the cover type or land use (LU), hydrologic soil group (HSG), and HC will result in a specific CN value based on the given tables. The CN table has to be developed based on LU, HSG and HC of the area. The TR-55 report<sup>238</sup> includes one of the widely used CN tables around the world.

The HC indicates the effects of the cover type and treatment on infiltration and runoff and is generally estimated from the density of plants and residue cover in the sample areas. Infiltration rates of soils vary considerably and are affected by subsurface permeability and surface intake rates. Soils are classified into four HSGs according to their minimum infiltration rate, which is obtained for bare soil after prolonged wetting<sup>238</sup>. In some study areas, the CN tables are defined based on the surface slope groups rather than the HC<sup>239–241</sup>.

Many studies have reported that  $\lambda = 0.2$  recommended by the USGS, obtained by analyzing small watersheds in the US, is not adequate in many scenarios. Therefore, in many regions, local CN tables and  $\lambda$  values are used for water assessments. Derived from recommended changes in the CN equation in recent studies<sup>226,242-249</sup>, Eq. 5 was proposed that expresses S as a function of the CN:

$$S = A * CN^x + B * CN^y + C * CN^z + D$$

Eq. 5

where A, B, C, D, x, y, and z can be any real value. For example, in the equation recommended by the USGS recommended, the result will be obtained using  $A = 25400$ ,  $x = -1$ ,  $D = -254$ , and  $B = C = y = z = 0$  (S in millimeters).

### **Antecedent Moisture Condition**

The AMCs in the soil in the basin is another important factor influencing the final CN value. AMCs are divided into the following three groups<sup>250</sup>:

AMCI: when the soil is almost dry.

AMCII: the average condition.

AMCIII: when basin soil is almost saturated from previous rainfall.

The CN values are assumed to express the average condition (AMCII). Depending on the cumulative rainfall of the last determined number of days (5 days is recommended by the USGS), the time of the year (either growing or dormant months), and the defined cumulative rainfall values for AMCI and AMCIII conditions, the CN values have to be corrected to correspond to the moisture condition. The following second-degree polynomial equation is proposed based on the existing literature for the AMC conversions<sup>251-253</sup>.

$$CN_{AMCI} = A * CN_{AMCII}^2 + B * CN_{AMCII} + C$$

Eq. 6

where  $CN_{AMCI}$  is the CN in AMCI or AMCIII;  $CN_{AMCII}$  is the original CN; and A, B, and C are any real number.

Note: in each study zone, the relationship between  $CN_{AMCI}$  and  $CN_{AMCIII}$  with  $CN_{AMCII}$  will potentially change in a new study area.

### **Connected Impervious Area and Unconnected Impervious Area**

The TR-55 report<sup>238</sup> recommendation for CCN calculation in urban areas was developed for CIAs and UIAs. CIAs are used when the runoff is directly connected to the drainage system, whereas UIAs are used when the runoff is spread over a pervious area as a sheet flow. The CIA and UIA CCN tables developed by the USGS are based on the assumption that impermeable areas have a CN of 98. The following equation (Eq. 7) was proposed to calculate the CCN in CIAs using an equation instead of Figs. 2 and 3 of the TR-55 report:

$$CCN = \frac{CIA}{100} * (98 - PCN) + PCN$$

Eq. 7

where PCN is the previous CN.

The CCN in the UIA is calculated using the previous formula if the total impervious area is less than 30%. In other cases, instead of Figs. 2–4 of the TR-55 report, Eq. 8 is proposed:

$$CCN = (3.3 - 1.7 \frac{UIAP}{TIAP}) * \frac{(120 - PCN)}{460} + PCN$$

Eq. 8

where TIAP and UIAP are the total and unconnected impervious area percentages, respectively.

### **4.2.3 Soil Water Reserve**

SWR or soil water storage is the maximum total amount of water that can be stored in the soil within the root zone of a plant. The SWR value depends on soil properties and can be calculated using the field capacity (FC), permanent wilting point (PWP), and root radial thickness (RRT) parameters. The formula to calculate SWR is as follows (Eq. 9) (Ministry of Agriculture, British Columbia, 2015):

$$SWR = (FC - PWP) * RRT$$

Eq. 9

where FC and PWP are volumetric water contents (dimensionless) and RRT and SWR are in millimeters (shown in light green in Fig. 1).

### **4.2.4 Potential Evapotranspiration**

Depending on the available data, time-step interval of the calculation, study area characteristics, and so forth, a suitable PET calculation method should be selected from a wide range of available methods. Generally, it is possible to classify the PET calculation methods in three main groups: temperature, radiation, and a combination of the two. To offer users with flexibility for different scenarios, they should

be able to choose from diverse PET calculation methods available in a tool. WaterpyBal provides the following PET methods: Blaney Criddle <sup>255</sup>, Hamon <sup>256</sup>, Linacre <sup>257</sup>, Romanenko <sup>258</sup>, Abtew <sup>259</sup>, Doorenbos–Pruitt <sup>260</sup>, Hargreaves <sup>261</sup>, Jensen and Haise <sup>262</sup>, Makkink <sup>263</sup>, McGuinness and Bordne <sup>264</sup>, Oudin <sup>265</sup>, Turc <sup>266</sup>, Kimberly Penman <sup>267</sup>, Penman <sup>268</sup>, FAO-56 <sup>269</sup>, Priestley and Taylor <sup>270</sup>, Penman-Monteith <sup>266</sup>, and Thom and Oliver <sup>271</sup>.

#### 4.2.5 Recharge, Real Evapotranspiration, and Deficit

The recharge (R), RET, and deficit (D) values can be calculated as follows using the PET, F, and SWR:

*If  $F > SWR - ASWR_{i-1} + PET$ :*

$$R = F - SWR - ASWR_{i-1} - PET$$

$$RET = PET$$

$$ASWR_i = SWR$$

$$D = 0$$

*Else:*

$$R = 0$$

*If  $PET > SWR - ASWR_{i-1}$ :*

$$RET = SWR - ASWR_{i-1}$$

$$ASWR_i = 0$$

$$D = SWR - ASWR_{i-1} - PET$$

*Else:*

$$RET = PET$$

$$ASWR_i = SWR - ASWR_{i-1} - PET$$

$$D = 0$$

where  $ASWR_{i-1}$  is the actual SWR from the previous time step, and  $ASWR_i$  is the actual SWR. An initial SWR value is needed to calculate the first step. Note: in an urban area, these parameters have to be recalculated to fit the urban zone characteristics.



#### 4.2.6 Urban Cycle

An urban water cycle, inspired by the literature, was designed <sup>272</sup>. Fig. 2 shows the following parameters that characterize the proposed urban cycle:

**Water Supply Network Consumption (WSNC)** is the amount of water supplied by the tap water network per area unit.

**Water Supply Network Loss (WSNL)** is the percentage of the WSNC loss.

**Direct Urban Evaporation (DUE)** is the percentage of water that evaporates from the sum of the rainfall and irrigation.

**Indirect Urban Evaporation (IUE)** is the percentage of water evaporating from the consumed water (e.g., washing clothes).

**Threshold of Rainfall per time step for Sewage loss (ThS)**. Usually the amount of sewage network loss differs based on the amount of water in the sewage network, which in turn depends on the rainfall events that are bigger than a certain threshold.

**Sewage Network Loss Normal (SNLN)** is the percentage of sewage network loss when the amount of rainfall per time step is lower than the defined threshold.

**Sewage Network Loss Rainy (SNLR)** is the percentage of sewage network loss when the amount of rainfall per time step is higher than the defined threshold.

**Runoff to Sewage (RtS)** is the percentage of runoff that finishes in the sewage network through catchments, and so forth.

**Direct Infiltration (DI)** is the percentage of the sum of the rainfall and irrigation that infiltrates directly into the ground.

**Water Consumption NOT from network (WCNN)** is the amount of consumed water supplied from sources other than the water supply network (e.g., wells).

**Water Consumption NOT from network loss (WCNNL)** is the percentage of loss of the aforementioned-consumed water.

**Water from Other Sources (WOS)** is water from other sources directed to the sewage network (e.g., underground infrastructure).

**Urban to Nonurban Area (UNUA)**: Each area unit in an urban zone can be a mix of urbanized and nonurbanized ambients. The SWB parameters such as infiltration, runoff, and PET in nonurbanized areas

can be calculated using the methods mentioned in Sections 2.2 and 2.4. These parameters can be recalculated for urbanized sections according to the specific water cycle of the urban areas. The urbanized to nonurbanized ratio of an urban area determines the mixing percentage of the water cycle parameters calculated using their respective methods. Appendix D shows the SWB calculations in urban areas using the aforementioned parameters.

Noteworthy, PET and infiltration values have to be recalculated before calculating the recharge, RET and deficit, taking into account the UNUA values. Then, the urban evaporation values have to be added to the RET using the determined UNUA proportion.

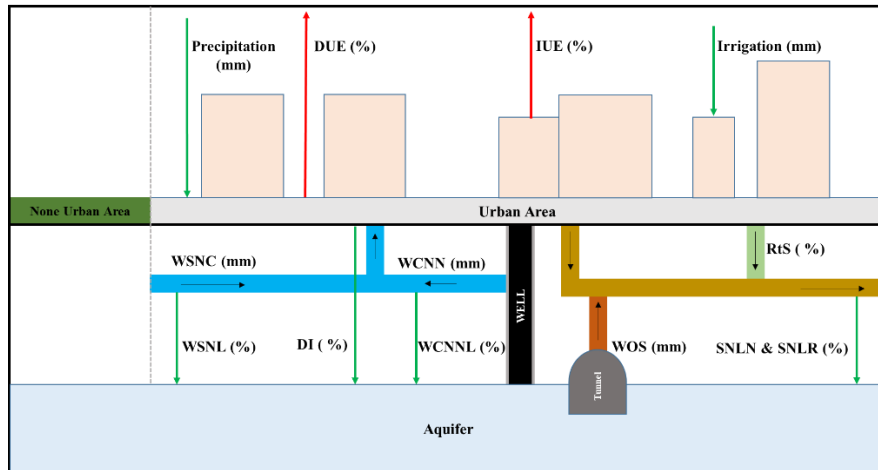


Figure 26 Simplified scheme of the urban water cycle. The abbreviations are explained in this section.

### 4.3 UNDER THE HOOD:

WaterpyBal utilizes the Network Common Data Form (netCDF) dataset because it is open-source, freely available, array-oriented, and portable. Many educational, research, and government projects utilize this dataset. netCDF can be accessed using many programming languages, computer programs, and geographic information system (GIS) software and allows the user to store, modify, and manage data from the study area for the desired number of variables and dimensions<sup>273</sup>.

WaterpyBal follows a modular design that aims to independently calculate each SWB parameter. As shown in Fig. 3, the general structure of WaterpyBal contains 9 classes and 17 methods. This structure allows easier application of WaterpyBal and facilitates further development of the library. WaterpyBal accepts a variety of inputs at each stage, as shown in Fig. 4. The following sections briefly describe the WaterpyBal classes:

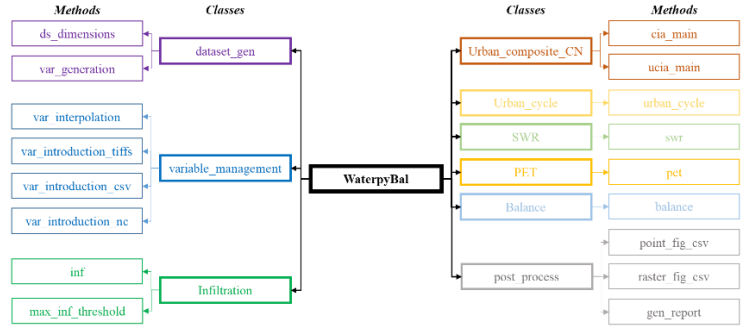


Figure 27 Simplified scheme of the main classes and methods of WaterpyBal. The colors used for each class correspond to the process with the same color in Fig. 1.

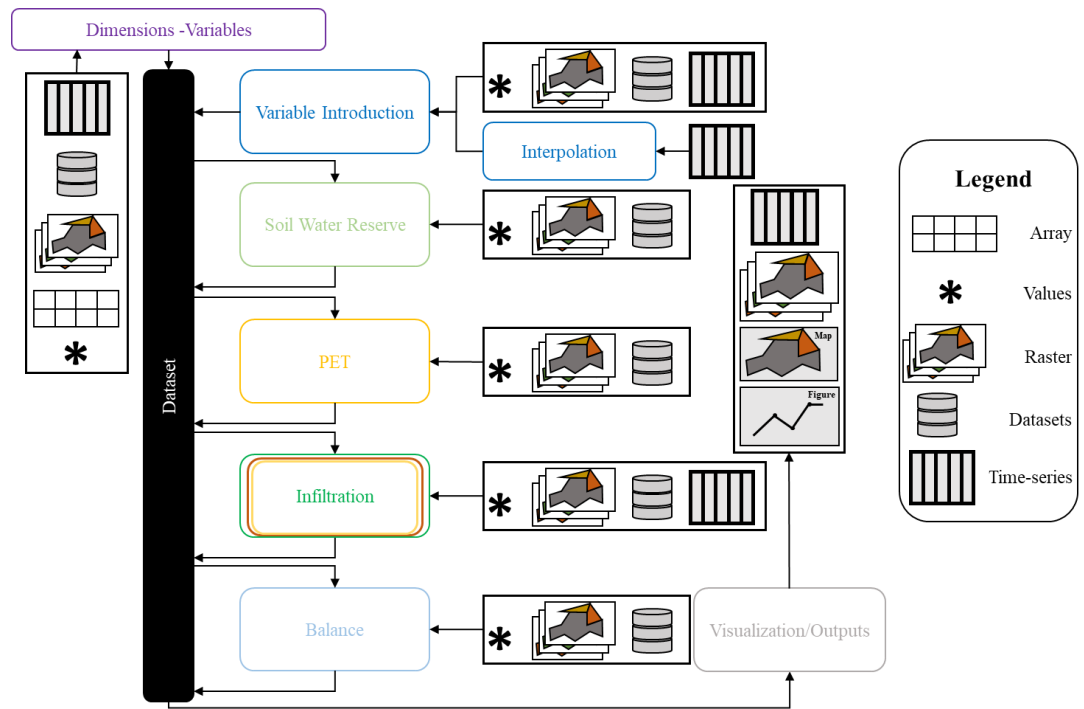


Figure 28 Scheme of the input types at each step of SWB calculation and modular structure of WaterpyBal. The colors used in each process correspond to the same stage with the same color in Figs. 1 and 3, respectively.

### 4.3.1 dataset\_gen

The *dataset\_gen* class comprises two methods (purple in Fig. 3). The *ds\_dimensions* method defines the spatial and temporal dimensions of the WaterpyBal netCDF dataset, whose boundary conditions can be determined by defining the limits manually, using an array, a datasheet containing all latitude and longitudes, a raster, or a netCDF dataset as a sample of the study area, as shown in Fig. 4. The time step of the dataset can be monthly, daily, or hourly and has to be defined at this stage.

The *var\_generation method* creates the WaterpyBal netCDF dataset and adds the user-defined additional variables. The spatial–temporal structure of the dataset and the availability of various open-source tools to manage it suggest that netCDF is a viable option to integrate the information available on the study area.

### **4.3.2 variable\_management**

The *variable\_management* class includes various methods to interpolate and/or store the available data (such as meteorological data) in the dataset (blue in Fig. 3). *var\_introduction\_nc*, *var\_introduction\_csv*, and *var\_introduction\_tiffs* allow the direct introduction of data from a netCDF, a.csv file, or a directory containing geotiffs data of the desired time steps to the WaterpyBal netCDF dataset, respectively.

The *var\_interpolation* method uses archives that contain time-series data in specific points of the study area and interpolates the variable at each time step for the whole study area using the nearest-neighbor, inverse-distance, moving average, linear, or a mix of the nearest-neighbor and inverse-distance methods.

### **4.3.3 SWR**

The *swr* method calculates the SWR (light green in Fig. 3). The soil characteristics can change with time, which results in temporal variability of the SWR. This method calculates the SWR using constant values, multiband rasters, or dataset variables. The user can determine if the SWR is constant in the study period, varies with each time step, or soil characteristics in the study area change at specific time steps. This will avoid unnecessary calculations and improve the execution time.

### **4.3.4 PET**

The *pet* method calculates the evapotranspiration via numerous methods (orange in Fig. 3). This method incorporates the PET methods using the pyet library <sup>274</sup> in a three-dimensional space. The needed variables for each method can be defined as a constant value using a raster (spatial variability) or variables that are already in the dataset (spatial–temporal variability). Section 2.4 shows a list of available PET methods.

### **4.3.5 Infiltration, Urban\_composite\_CN, and Urban\_cycle**

The *inf* method calculates the infiltration values using the CN-related methods, as mentioned in Section 2.2 (dark green in Fig. 3). The LU, HSG, and HC/altitude values of the study area can be introduced into WaterpyBal using different types of inputs, as shown in Fig. 4.

The predefined value for the initial abstraction values was 0.2. Because numerous studies suggest the variability of the initial abstraction value and CN formula, the advanced CN option in this method allows

the user to define S (potential maximum retention) by defining the A, B, C, x, y, and z values mentioned in Eq. 5.

The AMC can be considered in the *inf* method based on the five-day antecedent rainfall in the three AMC groups. The dormant and growing months can be distinguished using different thresholds to specify the final AMC group. The parameters of the second-degree polynomial equation between the AMCII and AMC<sub>i</sub> can be modified to allow flexibility in AMC calculations (Eq. 6).

The *max\_inf\_threshold* method can be used to force the maximum infiltration value in each pixel.

If the study area contains an urban zone, it is possible to calculate the CIA and UIA CCN values using the *CIA* and *UIA* methods, respectively (red in Fig. 3). The formula to calculate the CCN values is described in Section 2.2.3.

The *urban\_cycle* method calculates the urban cycle of the urban zone (yellow in Fig. 3). The urban cycle is characterized by 13 variables, as described in Section 2.6. These variables can be defined by constant values, raster, or a dataset.

Fig. 4 shows the input types that can be used in each aforementioned classes.

#### **4.3.6 Balance**

The *balance* method calculates the R, D, RET, and ASWR, as mentioned in Section 2.5. To calculate the mentioned variables, the initial water capacity must be determined as a starting point to calculate the ASWC in the following iterations. The initial ASWR can be defined as a percentage of the maximum water capacity or can be introduced as a two-dimensional (2D) raster or an array, as shown in Fig. 4.

#### **4.3.7 post\_process**

The *post\_process* class could be used to access the variables (measured and calculated) of the WaterpyBal dataset using 3 different methods and may generate diverse outputs, as shown in Fig. 4. The *point\_fig\_csv* method creates the time-series figures or datasheets in a specified point and time period. The *raster\_fig\_csv* method creates geotiffs, maps, or datasheets in the whole study area in a specified time period.

The *gen\_report* method generates a document that includes the calculated SWB variables in a specified area and/or the whole study area. Fig. F.1 in Appendix F shows some of the *post\_process* class outputs.

### **4.4 THE WORKFLOW**

The general steps to calculate the SWB using WaterpyBal and WaterpyBal Studio are shown in Fig. 4 from top to bottom and are as follows:

1. Defining the spatial–temporal properties of the WaterpyBal dataset.
2. Defining the additional variables that need to be included in the dataset. These additional variables can be the ones needed to calculate the PET and have to be introduced based on the PET method that will be used in the following steps.
3. Introducing the available/measured data to the database. This can be done using values, other netCDF databases, or from the time series of the points with measured data. If the time series of the points were introduced, WaterpyBal can interpolate data in each time step using the introduced time series for each variable using five different methods. The interpolation method is powered by the gdal Python library <sup>275</sup>. If the time interval of the input time series and WaterpyBal dataset are not the same, WaterpyBal will distribute or integrate the measured values after interpolation based on their nature to match the dataset time interval and store the values in the dataset.
4. Adding data to the database if the raster archives are available in all or some time steps for a variable.
5. Calculating SWR using the FC, PWP, and RRT.
6. Calculating PET using 1 of the 18 methods available in WaterpyBal.
7. Calculating infiltration. Infiltration can be calculated using the CN if the WaterpyBal dataset time interval is daily. If not, the infiltration has to be introduced directly in step 3. A multiband raster or separate raster archives has to be used to introduce the data needed to calculate the CN, containing the LU, HSG, and HC/Altitude (in case the CN table is based on the slope groups). The CCN can be calculated for urban areas.
8. Calculating urban cycle parameters based on the inputs mentioned in Section 6.2.
9. Calculating the SWB using the initial SWR.
10. Obtaining result reports and visualization of the outputs. WaterpyBal outputs can be divided into three different categories:
  1. Coordinate-based:

A coordinate can be specified, and a figure or datasheet of the time series in the determined time window will be generated for specified variables.
  2. Whole study area:

Maps or raster files of the study area in a determined time window for each time series for specified variables will be generated.

3. Region-based:

By introducing a raster that includes different regions of the study area, the SWB-related variables can be generated for each specified region. If the region raster is not provided to WaterpyBal, a report for the whole study area will be generated.

## **4.5 GRAPHIC USER INTERFACE OF THE WATERPYBAL LIBRARY**

The WaterpyBal Studio is the graphic user interface of WaterpyBal, which facilitates the use of WaterpyBal without any knowledge of programming languages. The WaterpyBal Studio covers most of the WaterpyBal library capabilities. Figs. F.2 and F.3 in Appendix F show the main page of the WaterpyBal Studio. More information is provided in Section 8.

## **4.6 SYNTHETIC EXAMPLE OF THE WATERPYBAL APPLICATION**

Here, this example is included to demonstrate the functionality of WaterpyBal and its results have been compared to four empirical methods <sup>276-279</sup> to show the relative coherency of the WaterpyBal results. The details of the empirical methods are given in Appendix E.

### **4.6.1 Designed Study Areas and Available Data**

The SWB was calculated in a synthetic study area of 30.58 km<sup>2</sup>, rasterized as 430 and 760 pixels in width and height, respectively, with a pixel size of 15 × 15 m. The study area comprises six different regions, as shown in Fig. 5. Table 1 shows the soil characteristics of each area. The second and third regions were urban areas with urban cycle parameters, as shown in Table 2.

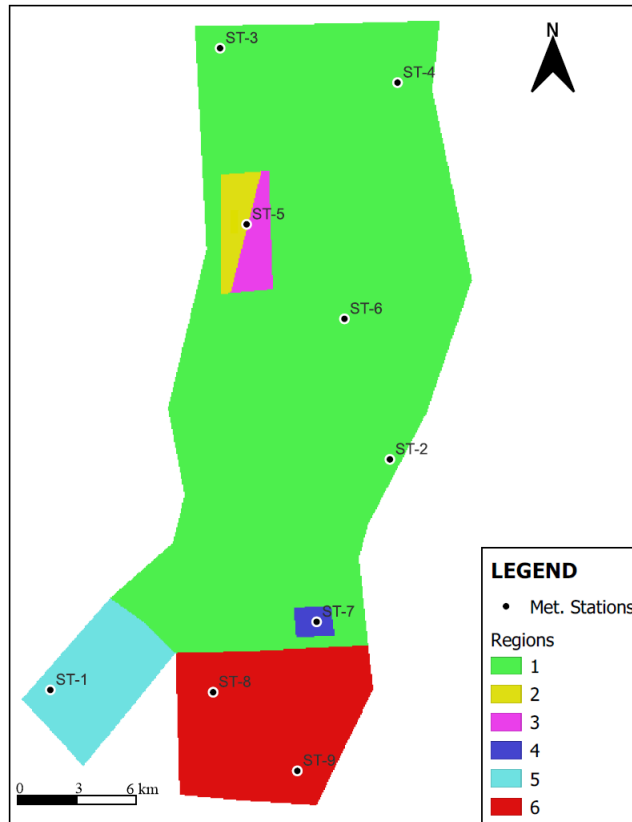


Figure 29 Synthetic example of a study area. The study area is divided into six regions with diverse characteristics. Regions 2 and 3 are urban areas. Nine meteorological stations have been shown as points in the maps.

Nine meteorological stations were included in the study area, as shown in Fig. 5. The available time series measured in these stations were daily precipitation (P) and monthly maximum, minimum, and average temperatures in 4 consecutive years (from 01/01/2018 to 31/12/2021). Fig. 6a shows the monthly average P of the nine stations, and Fig. 6b shows the average temperature measured at each meteorological station. The SWB was calculated for 4 years in daily time intervals, which is equal to 1461 time steps. In this example,  $\lambda = 0.2$  and the standard USGS CN table were used. The Hargreaves method was used in the study area to calculate the PET. The initial SWR was assumed 100%.



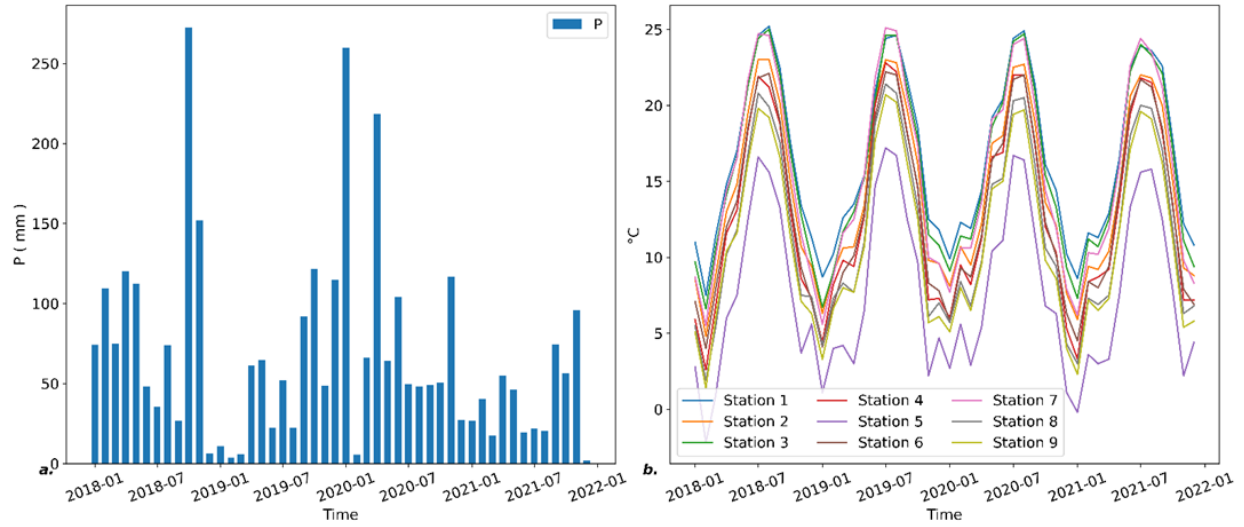


Figure 30 (a) Monthly average precipitation of nine meteorological stations. (b) Average temperature measurements for nine meteorological stations.

Table 6 Soil characteristics in six regions of the synthetic example. Zones 2 and 3, marked by \*, are urban zones. The same soil parameters have been designated to these two zones to demonstrate the variability of the urban cycle parameters in urban areas in the SWB.

Regions	FC	PWP	RRT	Area (Km2)	pixels	CN
1	0.37	0.25	0.20	23.14	102850	74
2*	0.23	0.12	0.20	0.48	2137	81
3*	0.23	0.12	0.20	0.48	2119	81
4	0.31	0.19	0.10	0.18	810	91
5	0.22	0.10	0.30	2.02	8964	66
6	0.38	0.25	0.35	4.28	19020	51

The urban area variables can be shown in Table 2. Since WaterpyBal accepts millimeter units for all inputs, including the urban parameters, the water consumption has to be converted to millimeters. The conversion is done by multiplying the water consumption per capita per day ( $m^3$ ) (first row in Table 2) and population (second row in Table 2), divided by the area ( $m^2$ ) (third row in Table 2), multiplied by 1000, which results in total water consumption value in mm (10.15 and 9.15 mm of total water consumption in region 2 and 3 respectively). It is assumed that to fulfill the total water consumption, 8 and 5 mm is from water supply network and 2.15 and 4.15 mm is from wells in regions 2 and 3, respectively. Section 8 provides access to the Jupyter notebook to reproduce the example.

Table 7 Urban cycle parameters in regions 2 and 3.

Parameter	Unit	Region 2	Region 3
<i>water consumption per capita per day</i>	$m^3$	0.27	0.28
<i>population</i>	-	2008	1731
<i>Area</i>	$m^2/pixels$	480825/2137	476775/2119
<i>Total water consumption</i>	mm	10.15	9.15
WSNC	mm	8	5
WSNL	%	7	10
DUE	%	14	18
IUE	%	4	6
SNLN	%	3	3
SNLR	%	8	8
ThS	mm	15	15
RtS	%	35	27
DI	%	25	39
WCNN	mm	2.15	4.15
WCNNL	%	6	7
WOS	mm	6.3	3
UNUA	%	100	60

#### 4.6.2 SWB Input Data Preparation:

Various raster files for WaterpyBal were prepared using QGIS software. All raster files had the same coordination system and pixel resolution as the study area. A six-band raster file containing the FC, PWP, RRT, LU, HSG, and HC was generated using the parameters shown in Table 1. A raster with the pixel values that identifies each region was used to create a final SWB report separated by the region numbers. The information in Table 2 was introduced to WaterpyBal using 13 raster files that characterized the urban area parameters in regions 2 and 3. The daily rainfall and monthly temperature time series of all nine meteorological stations were converted into two files (one for the daily rainfall and another one for the monthly temperature measurements) in a comma-separated value format. Section 8 provides all input files to reproduce the example.

#### 4.6.3 WaterpyBal Application:

The steps described in Section 4 were followed to obtain the SWB:

1. The WaterpyBal dataset was created for a time period ranging from 01/01/2018 to 31/12/2021 using a raster as a sample of the study area extent.
2. Additional variables, namely, minimum, maximum, and average temperatures, were included in the dataset to calculate the PET using the Hargreaves method.
3. The time series of measurements from nine meteorological stations for daily precipitation and monthly temperatures was interpolated in each time step via the linear method using WaterpyBal.

4. The SWR was calculated using the FC, PWP, and RRT bands of the raster file.
5. PET was calculated using the temperature data already interpolated and saved in the dataset using WaterpyBal in step 3.
6. Using the WaterpyBal default USGS standard table, the CNs were identified using the LU, HSG, and HC from the multiband raster. WaterpyBal default AMC correction was implemented.
7. Urban zone raster files were introduced into WaterpyBal.
8. The SWB was calculated.
9. The result was visualized and exported, as explained in Section 6.4.

#### 4.6.4 Example Results and Discussion:

The complete output files can be reproduced and found, as described in Section 8. Here, some relevant results are discussed. Fig. 7 shows the P, R, and RET in regions 1 and 4. To evaluate the coherency of the results, the infiltration data calculated by WaterpyBal were introduced to the EASY BAL application in nonurban regions (1, 4, 5, and 6), which resulted in less than 0.2% difference between the average annual R values. Moreover, four empirical methods were used to calculate the annual R. Appendix E provides the details of the four empirical methods. The empirical methods for regions 2 and 3 were not included since these methods are not designed for urban regions.

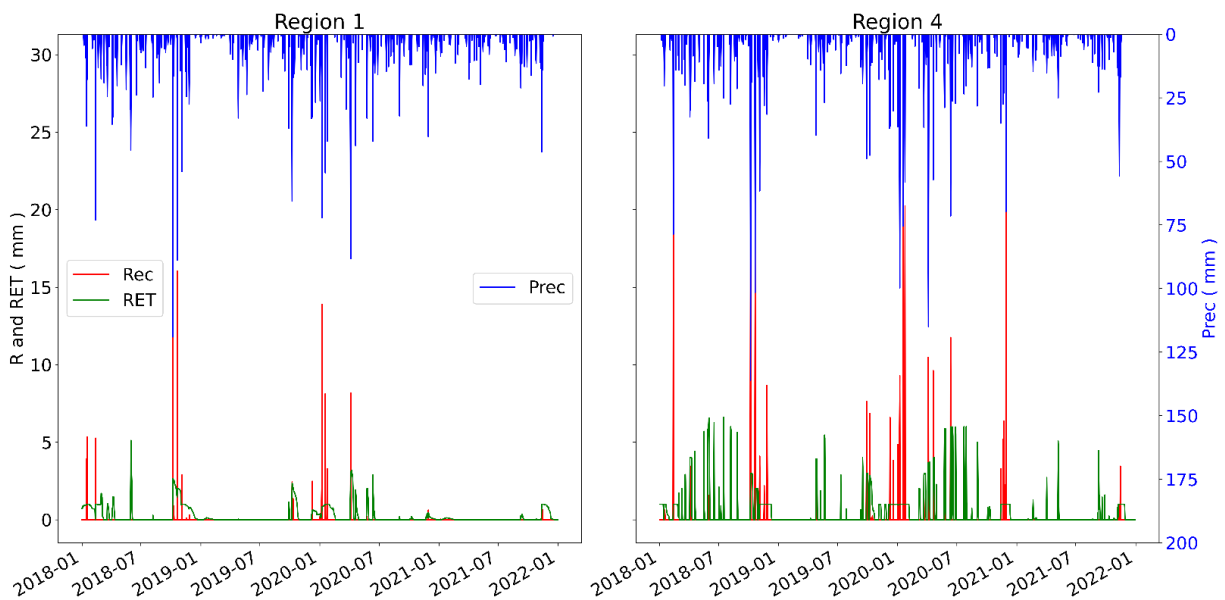


Figure 31 Precipitation, recharge, and RET values in regions 1 and 4. Values in mm/day.

Table 3 shows the minimum, maximum, standard deviation, and average of R for four empirical methods in each year of each zone compared to the R calculated using WaterpyBal. The R calculated using WaterpyBal is calculated in daily time steps and integrated to annual R to be comparable to the empirical methods. In Fig. 8, the dashed lines show the annual recharge calculated by WaterpyBal and the vertical bars show the two standard deviations range from the average of the empirical methods. Fig. 8a in 2019,

Fig. 8c in 2018 and 2019, Fig. 8d in 2019 and Table 3 show that the R calculated by WaterpyBal is coherent with the empirical methods, except in scenarios with a low P where WaterpyBal estimates a lower R than the empirical methods. This is because the R calculated by empirical methods is a direct function of the cumulative annual P and not sensitive to the P distribution throughout the year.

Table 8 Annual recharge calculated by WaterpyBal and the statistics of the annual recharge calculated by four empirical methods. Values in mm/year. Bold fonts show the year and regions that WaterpyBal estimation is outside the maximum and minimum range that is determined by empirical methods.

Region	Year	WaterpyBal annual recharge estimation	Empirical methods statistics			
			Minimum	Maximum	Standard deviation	Mean
1	2018	65	56	163	48	75
	<b>2019</b>	<b>6</b>	<b>19</b>	<b>56</b>	<b>16</b>	<b>33</b>
	2020	38	53	131	35	65
	2021	1	0	42	21	26
4	2018	93	56	168	50	88
	2019	29	22	57	15	42
	2020	142	60	221	73	105
	2021	4	0	44	21	33
5	<b>2018</b>	<b>5</b>	<b>53</b>	<b>125</b>	<b>32</b>	<b>53</b>
	2019	0	0	50	23	22
	<b>2020</b>	<b>1</b>	<b>48</b>	<b>74</b>	<b>13</b>	<b>39</b>
	2021	0	0	37	18	16
6	<b>2018</b>	<b>18</b>	<b>60</b>	<b>226</b>	<b>75</b>	<b>53</b>
	<b>2019</b>	<b>0</b>	<b>46</b>	<b>66</b>	<b>9</b>	<b>29</b>
	<b>2020</b>	<b>19</b>	<b>59</b>	<b>212</b>	<b>69</b>	<b>52</b>
	2021	0	0	44	21	21

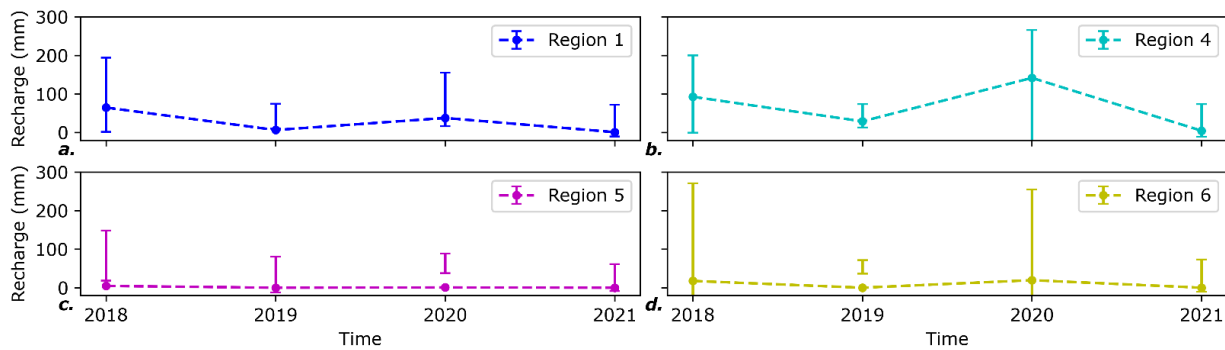


Figure 32 Dashed lines show the annual recharge calculated by WaterpyBal, and the vertical bars show the two standard deviations of the average of the empirical methods. Negative values are considered 0 in empirical models. a, b, c, and d correspond to the regions 1, 4, 5, and 6, respectively. Values in mm/year.

Figs. 9 a and b show the daily P, R, and evaporation (Ep) in urban regions. Ep in the two regions follows a distinct pattern because the UNUA is defined as 100% and 60% in regions 2 and 3, respectively. The Ep in region 2 is solely the result of the urban cycle, whereas the Ep in region 3 is the ratio of the RET and urban cycle Ep.

The average R value in regions 2 and 3 is 1.9 and 0.8 mm per day, respectively, for the 4-year period. The minimum R value produced by just by urban activities in the absence of precipitation (e.g., network loss and sewage leakage) for regions 2 and 3 is 1.15 and 0.67 mm per day, respectively, for the 4-year period.

To see the effects of urbanization on recharge, the SWB was calculated in an urban area of 1 m<sup>2</sup> with the same urban cycle parameters as region 2, changing the UNUA parameter from 0% to 100%. Fig. 9c. shows the average daily R for the 4-year period calculated with different UNUA values. Fig. 9c. show that by increasing the ratio of the nonurban areas, R increases, which highlights the importance of green areas in urban regions for recharge and flood management.

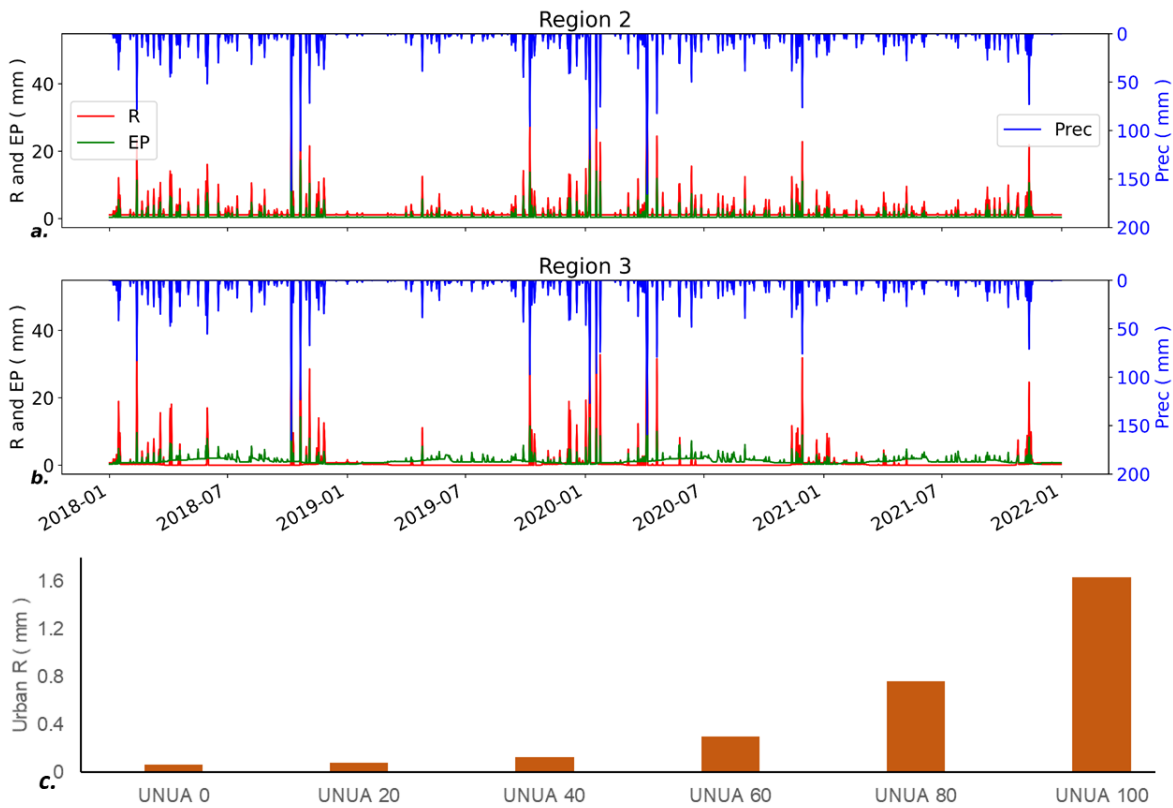


Figure 33 (a, b) Precipitation, recharge, and evaporation in regions 2 and 3, respectively. (c) Recharge evolution by changing the UNUA value in an urban area. Values in mm/day.

## 4.7 CONCLUSIONS

WaterpyBal is an open-source Python library dedicated to SWB calculations. It has hourly, daily, and monthly temporal variability and accepts a wide variety of data types as the inputs of the model. WaterpyBal unifies different stages of SWB calculation into an integrated library: from the spatial interpolation of the input parameters to creating reports, figures, and maps of the results. WaterpyBal utilizes netCDF, which is a well-known dataset used in many other available tools. The modular algorithm structure of WaterpyBal

allows the easy development and application of WaterpyBal in the future. A synthetic example was used to demonstrate the ease of use and coherency of WaterpyBal.

Moreover, the WaterpyBal Studio is a friendly graphic user interface of WaterpyBal containing the commonly used features of WaterpyBal. It is accompanied by a guide at each stage to clarify each available option.

The extensive urban cycle takes into account different crucial parameters of the SWB and recharge quantification in an urban environment, which results in a more realistic urban model.

Nevertheless, WaterpyBal has some limitations. It should not be directly used in study areas where the horizontal water flow is prominent since it just takes into account the vertical water flow. Furthermore, calculating infiltration using WaterpyBal is limited to daily time intervals since the CN method is used for infiltration calculation. Infiltration has to be introduced to WaterpyBal as an input parameter if there are different time steps.

The easily generated reports, maps, and figures of the introduced and resulting variables save the user analysis time, allowing faster result evaluation. Overall, WaterpyBal can help in sustainable groundwater management in urban and nonurban areas.

## **4.8 SOFTWARE AND DATA AVAILABILITY**

WaterpyBal is available to download freely in <https://github.com/IDAEA-EVS/waterpybal> Under AGPL-3.0 License.

Year first available: 2023

Programming language: Python

Dependencies: numpy, pandas, xarray, matplotlib, netCDF4, osgeo, rasterio, richdem, rioxarray, pyet

Developed by Ashkan Hassanzadeh

Contact information: [ashkan.hassanzadeh@csic.es](mailto:ashkan.hassanzadeh@csic.es).

Refer to <https://readthedocs.org/projects/waterpybal> for additional information about the installation, default values of the arguments, the explanation and the usage.

The input data alongside the Jupyter notebook of the synthetic example is available at <https://github.com/IDAEA-EVS/waterpybal>.

WaterpyBal Studio installer, installation guide, user manual and inputs for the synthetic example are available to download freely at <https://doi.org/10.20350/digitalCSIC/15191>.

## **5 Conclusions**

Groundwater characterization and management, an essential aspect for life as we know it, requires the integration of large amounts of data. This can be eased with the use of numerical tools that automate certain calculations, which in turn, helps in avoiding biases and saves processing time. Geopropy, Isocompy and WaterpyBal are the three tools that are specifically created as part of this thesis that provide more accurate and reliable numerical models for aspects of groundwater administration. The new tools are all open source Python libraries, that deal with geological cross sections, water isotopes and soil water balances. The tools incorporate codes for visualization of the solutions obtained, in terms of cross sections or maps, which also eases the data interpretation.

- Geopropy is a library that utilizes a data-driven approach to generate 3D geological cross-sections from both surface and subsurface geological data. It aims to assist geologists in making consistent decisions by identifying zones that may have multiple outcomes in a cross-section. This feature can potentially help users to analyze different geological scenarios based on the available data. This tool does not replace implicit models but works as an intelligent agent to generate cross-sections explicitly. It has been validated by applying it to synthetic profiles, making it a valuable tool for geologists. The library can speed up the process of creating geological cross-sections, which can be time-consuming.

Geopropy identifies the uncertainty caused by complex structures or lack of data by detecting zones with more than one possible outcome. It facilitates the decision-making process in these zones for the user in three degrees of freedom. The first degree of freedom is automatic, where the algorithm generates cross-sections when available information results in a unique outcome. The second degree of freedom is semi-automatic, where the algorithm asks for decisions on how to complete the geological unit contacts if there are multiple outcomes. The third degree of freedom is manual, where the algorithm enters the manual stage to complete the cross-section if new geospatial information or more complex decisions are required. Geopropy's main advantage is its ability to combine the efficiency of implicit modeling with the accuracy and detail of explicit modeling. By automating parts of the workflow and incorporating expert opinion and data, Geopropy supports the decision-making process of geologists, allowing them to create more accurate and detailed geological models with greater efficiency.

- Isocompy is designed for regression-statistical modeling and analysis of natural water isotopic compositions. Its multistage procedure considers various potential features that can impact the water isotopic signature, including meteorological measurements, sea surface temperatures, and reanalysis-derived variables. The library simplifies and optimizes the water isotopic interpretation



in complex areas with limited and non-uniform data. Isocompy can deliver isotopic spatial distribution and the estimated local meteoric water line data, and the feature selection procedure allows for a quick yet extensive study of the influencing factors of precipitation's isotopic composition. Additionally, Isocompy generates extensive outputs, making the evaluation process user-friendly. The library is flexible and adaptable, depending on the amount of data available in time and space, and has the capacity to apply various regression methods. The functionality of this tool has been validated by implementing it on the meteorological features and isotopic composition of precipitation in N Chile.

- WaterpyBal calculates soil water balances. It incorporates different stages of water balance modelling such as spatial data interpolation, evaporation, evapotranspiration and infiltration calculation, taking into account the soil characteristics and urban water cycle parameters and post-processing using a widely known dataset. Its modular algorithm structure makes it easy to develop and enhance the tool in the future. The extensive urban cycle in WaterpyBal accounts for various critical parameters of soil water balance and recharge quantification in urban environments, resulting in a more realistic urban model. The tool generates extensive outputs that aid in a quick result evaluation. Moreover, WaterpyBal Studio is a graphical user interface that contains frequently used features of WaterpyBal. A synthetic example is provided to demonstrate the tool's user-friendliness and coherence.

Developing these tools in a widely-used programming language, opening the source codes and modular algorithm designs ensure the possibility of a wide use and further development of the libraries and contribute to the reproducible research. Moreover, they can also be coupled to other modern existing decision-making libraries. These tools would aid the user in the different stages of conceptual and numerical groundwater modelling and facilitate groundwater management.

## References

1. Liu, J., Fu, Z. & Liu, W. Impacts of precipitation variations on agricultural water scarcity under historical and future climate change. *J. Hydrol.* **617**, 128999 (2023).
2. Zhang, C. Y. & Oki, T. Water pricing reform for sustainable water resources management in China's agricultural sector. *Agric. Water Manag.* **275**, 108045 (2023).
3. Criollo, R. *et al.* AkvaGIS: An open source tool for water quantity and quality management. *Comput. Geosci.* **127**, 123–132 (2019).
4. Kifanyi, G. E., Ndambuki, J. M., Odai, S. N. & Gyamfi, C. Quantitative management of groundwater resources in regional aquifers under uncertainty: A retrospective optimization approach. *Groundw. Sustain. Dev.* **8**, 530–540 (2019).
5. Elshall, A. S. *et al.* Groundwater sustainability: a review of the interactions between science and policy. *Environ. Res. Lett.* **15**, 093004 (2020).
6. Enemark, T., Peeters, L. J. M., Mallants, D. & Batelaan, O. Hydrogeological conceptual model building and testing: A review. *J. Hydrol.* **569**, 310–329 (2019).
7. Glass, J., Junghanns, R., Schlick, R. & Stefan, C. The INOWAS platform: A web-based numerical groundwater modelling approach for groundwater management applications. *Environ. Model. Softw.* **155**, 105452 (2022).
8. Sikdar, P. K. Numerical groundwater modelling. *Groundw. Dev. Manag. Issues Challenges South Asia* 191–207 (2018). doi:10.1007/978-3-319-75115-3\_7/FIGURES/4
9. Refsgaard, J. C. *et al.* Review of strategies for handling geological uncertainty in groundwater flow and transport modeling. *Adv. Water Resour.* **36**, 36–50 (2012).
10. Randle, C. H., Bond, C. E., Lark, R. M. & Monaghan, A. A. Can uncertainty in geological cross-section interpretations be quantified and predicted? *Geosphere* **14**, 1087–1100 (2018).
11. Kumar, T. S. S. Resource Modelling of Iron Ore Deposit using Surpac Software. *J. Geol. Soc. India* **97**, 559–559 (2021).
12. Bowden, R. A. Building confidence in geological models. *Geol. Soc. London, Spec. Publ.* **239**, 157–173 (2004).
13. Randle, C. H., Bond, C. E., Lark, R. M. & Monaghan, A. A. *Uncertainty in geological*

- interpretations: Effectiveness of expert elicitations. Geosphere* **15**, 108–118 (Geological Society of America, 2019).
14. Jafari, T., Kiem, A. S., Javadi, S., Nakamura, T. & Nishida, K. Using insights from water isotopes to improve simulation of surface water-groundwater interactions. *Sci. Total Environ.* **798**, 149253 (2021).
  15. Custodio, E. *Recarga natural a los acuíferos, metodología y soporte de la isotopía del agua : aplicación a la planificación hidrológica y conocimiento de las aguas subterráneas en España : informe RAEMIA.* (UPC, 2019).
  16. VALDIVIELSO, S., HASSANZADEH, A., VÁZQUEZ-SUÑÉ, E. & CUSTODIO, E. The factors that condition the isotopic composition of rain in northern Chile. in *Goldschmidt 3000* (2019).
  17. Putman, A. L., Fiorella, R. P., Bowen, G. J. & Cai, Z. A Global Perspective on Local Meteoric Water Lines: Meta-analytic Insight Into Fundamental Controls and Practical Constraints. *Water Resour. Res.* **55**, 6896–6910 (2019).
  18. Xi, X. A Review of Water Isotopes in Atmospheric General Circulation Models: Recent Advances and Future Prospects. (2014). doi:10.1155/2014/250920
  19. Steen-Larsen, H. C., Risi, C., Werner, M., Yoshimura, K. & Masson-Delmotte, V. Evaluating the skills of isotope-enabled general circulation models against in situ atmospheric water vapor isotope observations. *J. Geophys. Res. Atmos.* **122**, 246–263 (2017).
  20. Tsuchihara, T., Shirahata, K., Ishida, S. & Yoshimoto, S. Application of a Self-Organizing Map of Isotopic and Chemical Data for the Identification of Groundwater Recharge Sources in Nasunogahara Alluvial Fan, Japan. *Water* 2020, Vol. 12, Page 278 **12**, 278 (2020).
  21. Weatherl, R. K., Henao Salgado, M. J., Ramgraber, M., Moeck, C. & Schirmer, M. Estimating surface runoff and groundwater recharge in an urban catchment using a water balance approach. *Hydrogeol. J.* 2021 297 **29**, 2411–2428 (2021).
  22. Scanlon, B. R., Healy, R. W. & Cook, P. G. Choosing appropriate techniques for quantifying groundwater recharge. *Hydrogeol. J.* **10**, 18–39 (2002).
  23. Simmers, I., Hendrickx, J. M. H., Kruseman, G. E. & Rushton, K. R. *Recharge of phreatic aquifers in (Semi-)arid areas. Recharge of Phreatic Aquifers in (Semi-)Arid Areas* (CRC Press, 2017). doi:10.1201/9780203741191

24. Serrano-Juan, A. *et al.* Customization, extension and reuse of outdated hydrogeological software. (2020). doi:10.1344/GeologicaActa2020.18.9
25. Arnold, J. G., Srinivasan, R., Muttiah, R. S. & Williams, J. R. LARGE AREA HYDROLOGIC MODELING AND ASSESSMENT PART I: MODEL DEVELOPMENT1. *JAWRA J. Am. Water Resour. Assoc.* **34**, 73–89 (1998).
26. Touhami, I. *et al.* Comparative performance of soil water balance models in computing semi-arid aquifer recharge. <https://doi.org/10.1080/02626667.2013.802094> **59**, 193–203 (2013).
27. Dubois, E., Larocque, M., Gagné, S. & Meyzonnat, G. Simulation of long-term spatiotemporal variations in regional-scale groundwater recharge: Contributions of a water budget approach in cold and humid climates. *Hydrol. Earth Syst. Sci.* **25**, 6567–6589 (2021).
28. Samper, J. *et al.* Using hydrological models and Geographic Information Systems for water resources evaluation: GIS-VISUAL-BALAN and its application to Atlantic basins in Spain (Valiñas) and Portugal (Serra da Estrela). *IAHS-AISH Publ.* **310**, 259–266 (2007).
29. Batelaan, O. & Smedt, F. WetSpass: A flexible, GIS based, distributed recharge methodology for regional groundwater modelling. *IAHS-AISH Publ.* (2001).
30. Gochis, J. & Chen, F. *Hydrological Enhancements to the Community Noah Land Surface Model.* (2003). doi:10.5065/D60P0X00
31. Lin, Y. F., Wang, J. & Valocchi, A. J. PRO-GRADE: GIS toolkits for ground water recharge and discharge estimation. *Ground Water* **47**, 122–128 (2009).
32. Šimůnek, J., Genuchten, M. T. & Šejna, M. Recent Developments and Applications of the HYDRUS Computer Software Packages. *Vadose Zo. J.* **15**, 1–25 (2016).
33. Pomeroy, J. W. *et al.* The cold regions hydrological model: a platform for basing process representation and model structure on physical evidence. *Process* **21**, 2650–2667 (2007).
34. Yang, B., Yuan, S., Liang, Y. & Liu, J. Investigation of overburden failure characteristics due to combined mining: case study, Henan Province, China. *Environ. Earth Sci.* **80**, (2021).
35. Hawie, N., Al-Wazzan, H., Al-Ali, S. & Al-Sahlan, G. De-risking hydrocarbon exploration in lower Jurassic carbonate systems of Kuwait through forward stratigraphic models. *Mar. Pet. Geol.* **123**, 104700 (2021).
36. Kerrou, J., Deman, G., Tacher, L., Benabderrahmane, H. & Perrochet, P. Numerical and polynomial

- modelling to assess environmental and hydraulic impacts of the future geological radwaste repository in Meuse site (France). *Environ. Model. Softw.* **97**, 157–170 (2017).
37. Pasculli, A., Palermi, S., Sarra, A., Piacentini, T. & Miccadei, E. A modelling methodology for the analysis of radon potential based on environmental geology and geographically weighted regression. *Environ. Model. Softw.* **54**, 165–181 (2014).
  38. Hillier, M. J., Schetselaar, E. M., de Kemp, E. A. & Perron, G. Three-Dimensional Modelling of Geological Surfaces Using Generalized Interpolation with Radial Basis Functions. *Math. Geosci.* **46**, 931–953 (2014).
  39. Lajaunie, C., Courrioux, G. & Manuel, L. Foliation fields and 3D cartography in geology: Principles of a method based on potential interpolation. *Math. Geol.* **29**, 571–584 (1997).
  40. Calcagno, P., Chilès, J. P., Courrioux, G. & Guillen, A. Geological modelling from field data and geological knowledge. Part I. Modelling method coupling 3D potential-field interpolation and geological rules. *Phys. Earth Planet. Inter.* **171**, 147–157 (2008).
  41. Gonçalves, Í. G., Kumaira, S. & Guadagnin, F. A machine learning approach to the potential-field method for implicit modeling of geological structures. *Comput. Geosci.* **103**, 173–182 (2017).
  42. Souche, L., Lepage, F. & Iskenova, G. Volume based modeling - Automated construction of complex structural models. in *75th European Association of Geoscientists and Engineers Conference and Exhibition 2013 Incorporating SPE EUROPEC 2013: Changing Frontiers* 5033–5037 (2013). doi:10.3997/2214-4609.20130037
  43. Iskenova, G. *et al.* A Novel Workflow for Modelling Complex Compartmentalised Structures Leads to Enhanced Field Development Strategy. (2016). doi:10.4043/26643-ms
  44. Brandes, C. & Tanner, D. C. Fault-related folding: A review of kinematic models and their application. *Earth-Science Reviews* **138**, 352–370 (2014).
  45. Caumon, G. Towards stochastic time-varying geological modeling. *Math. Geosci.* **42**, 555–569 (2010).
  46. Lemon, A. M. & Jones, N. L. Building solid models from boreholes and user-defined cross-sections. *Comput. Geosci.* **29**, 547–555 (2003).
  47. Muzik, J., Vondráčková, T., Sitányiová, D., Plachý, J. & Nývlt, V. Creation of 3D Geological Models Using Interpolation Methods for Numerical Modelling. *Procedia Earth Planet. Sci.* **15**, 25–

- 30 (2015).
48. Jessell, M. W. & Valenta, R. K. Structural geophysics: Integrated structural and geophysical modelling. *Comput. Methods Geosci.* **15**, 303–324 (1996).
  49. Florian Wellmann, J., Thiele, S. T., Lindsay, M. D. & Jessell, M. W. Pynoddy 1.0: An experimental platform for automated 3-D kinematic and potential field modelling. *Geosci. Model Dev.* **9**, 1019–1035 (2016).
  50. De La Varga, M., Schaaf, A. & Wellmann, F. GemPy 1.0: Open-source stochastic geological modeling and inversion. *Geosci. Model Dev.* **12**, 1–32 (2019).
  51. Schaaf, A., de la Varga, M., Wellmann, F. & Bond, C. Constraining stochastic 3-D structural geological models with topology information using Approximate Bayesian Computation using GemPy 2.1. *Geosci. Model Dev. Discuss.* 1–24 (2020). doi:10.5194/gmd-2020-136
  52. BRGM. GDM Suite. (2020). Available at: <https://www.brgm.fr/en/software/gdm-suite-software-suite-allowing-model-represent-visualize-geoscientific-data-geological>. (Accessed: 21st August 2021)
  53. Petroleum Experts. MOVE. (2019). Available at: <https://www.petex.com/products/move-suite/>. (Accessed: 21st August 2021)
  54. Maptek. Eureka. (2021). Available at: <https://www.maptek.com/products/eureka/>. (Accessed: 21st August 2021)
  55. Seequent. Leapfrog. (2021). Available at: <https://www.seequent.com/products-solutions/leapfrog-geo/>. (Accessed: 21st August 2021)
  56. Seequent. Geosoft | Oasis montaj. (2021). Available at: <https://www.seequent.com/products-solutions/geosoft-oasis-montaj/>. (Accessed: 21st August 2021)
  57. Velasco, V. *et al.* The use of GIS-based 3D geological tools to improve hydrogeological models of sedimentary media in an urban environment. *Environ. Earth Sci.* **68**, 2145–2162 (2013).
  58. Alcaraz, M., Vázquez-Suñé, E., Velasco, V. & Diviu, M. 3D GIS-based visualisation of geological, hydrogeological, hydrogeochemical and geothermal models. *Zeitschrift der Dtsch. Gesellschaft für Geowissenschaften* **167**, 377–388 (2016).
  59. Källgården, J. & Spångmyr, H. QGIS Python Plugins Repository. (2020). Available at: <https://plugins.qgis.org/plugins/midvatten/>. (Accessed: 21st August 2021)

60. GRASS Development Team. GRASS GIS. (2020). Available at: <https://grass.osgeo.org/>. (Accessed: 21st August 2021)
61. Serrano, R. Geomodelr's documentation. (2019). Available at: <https://geomodelr.readthedocs.io/en/latest/>. (Accessed: 21st August 2021)
62. RockWare. RockWorks. (2020). Available at: <https://www.rockware.com/product/rockworks/>. (Accessed: 21st August 2021)
63. Dassault Systèmes. GEOVIA Surpac. (2021). Available at: <https://www.3ds.com/products-services/geovia/products/surpac/>. (Accessed: 21st August 2021)
64. Datamine. Datamine. (2021). Available at: <https://www.dataminesoftware.com/>. (Accessed: 21st August 2021)
65. I-GIS. GeoScene3D | 3D Modelling | GeoScene. (2020). Available at: <https://geoscene3d.com/>. (Accessed: 21st August 2021)
66. British Geological Survey. Groundhog Desktop. (2020). Available at: <https://www.bgs.ac.uk/technologies/software/groundhog/>. (Accessed: 21st August 2021)
67. Cullen, H., Kessler, H., Wood, B. & Mathers, S. *Geological surveying and investigation in 3D*. (2010).
68. Bryant, I. D. & Flint, S. S. Quantitative Clastic Reservoir Geological Modelling: Problems and Perspectives. in *The Geological Modelling of Hydrocarbon Reservoirs and Outcrop Analogues 1–20* (Blackwell Publishing Ltd., 2009). doi:10.1002/9781444303957.ch1
69. Schlumberger. Petrel. (2015). Available at: <https://www.software.slb.com/products/petrel/>. (Accessed: 21st August 2021)
70. Intrepid Geophysics. GeoModeller. (2020). Available at: <https://www.intrepid-geophysics.com/product/geomodeller/>. (Accessed: 21st August 2021)
71. Paradigm. GOCAD. (2015). Available at: <https://www.pdgm.com/products/GOCAD>. (Accessed: 21st August 2021)
72. Abadi, M. *et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*. (2016).
73. Pedregosa, F. *et al. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

74. Paszke, A. *et al.* *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. *Advances in Neural Information Processing Systems* **32**, (2019).
75. Salvatier, J., Wiecki, T. V. & Fonnesbeck, C. Probabilistic programming in Python using PyMC3. *PeerJ Comput. Sci.* **2016**, e55 (2016).
76. Dynamic Graphics. EarthVision Software. (2009). Available at: <https://www.dgi.com/>. (Accessed: 24th August 2021)
77. Golden Software. Strater. (2021). Available at: <https://www.goldensoftware.com/products/strater>. (Accessed: 24th August 2021)
78. Idrees, M. B., Jehanzaib, M., Kim, D. & Kim, T.-W. Comprehensive evaluation of machine learning models for suspended sediment load inflow prediction in a reservoir. *Stoch. Environ. Res. Risk Assess.* (2021). doi:10.1007/s00477-021-01982-6
79. Chai, Z., Nwachukwu, A., Zagayevskiy, Y., Amini, S. & Madasu, S. An integrated closed-loop solution to assisted history matching and field optimization with machine learning techniques. *J. Pet. Sci. Eng.* **198**, (2021).
80. Chaikine, I. A. & Gates, I. D. A machine learning model for predicting multi-stage horizontal well production. *J. Pet. Sci. Eng.* **198**, (2021).
81. Hörning, S. & Haese, B. RMWSPy (v 1.1): A Python code for spatial simulation and inversion for environmental applications. *Environ. Model. Softw.* **138**, 104970 (2021).
82. Volk, J. M. & Turner, M. A. PRMS-Python: A Python framework for programmatic PRMS modeling and access to its data structures. *Environ. Model. Softw.* **114**, 152–165 (2019).
83. White, J. T., Hemmings, B., Fienen, M. N. & Knowling, M. J. Towards improved environmental modeling outcomes: Enabling low-cost access to high-dimensional, geostatistical-based decision-support analyses. *Environ. Model. Softw.* **139**, 105022 (2021).
84. Heron, M., Hanson, V. L. & Ricketts, I. Open source and accessibility: advantages and limitations. *J. Interact. Sci.* **2013 11 1**, 1–10 (2013).
85. Allen, D. M., Schuurman, N., Deshpande, A. & Scibek, J. Data integration and standardization in cross-border hydrogeological studies: A novel approach to hydrostratigraphic model development. *Environ. Geol.* **53**, 1441–1453 (2008).
86. Van Rossum, G. *Centrum voor Wiskunde en Informatica Python tutorial*. (1995).



87. Bond, C. E. Uncertainty in structural interpretation: Lessons to be learnt. *Journal of Structural Geology* (2015). doi:10.1016/j.jsg.2015.03.003
88. Jessell, M. W., Wellmann, J. F., Pakyuz-Charrier, E., Lindsay, M. & Thiele, S. T. The topology of geology 2: Topological uncertainty. *J. Struct. Geol.* **91**, 74–87 (2016).
89. Refsgaard, J. C., van der Sluijs, J. P., Højberg, A. L. & Vanrolleghem, P. A. Uncertainty in the environmental modelling process - A framework and guidance. *Environ. Model. Softw.* **22**, 1543–1556 (2007).
90. Wellmann, F. & Caumon, G. 3-D Structural geological models: Concepts, methods, and uncertainties. in *Advances in Geophysics* **59**, 1–121 (Elsevier, 2018).
91. Wellmann, J. F. & Regenauer-Lieb, K. Uncertainties have a meaning: Information entropy as a quality measure for 3-D geological models. *Tectonophysics* **526–529**, 207–216 (2012).
92. White, J. T., Fienen, M. N. & Doherty, J. E. A python framework for environmental model uncertainty analysis. *Environ. Model. Softw.* **85**, 217–228 (2016).
93. Davis, G. H., Reynolds, S. J. & Kluth, C. F. *Structural geology of rocks and regions*. (John Wiley & Sons, 2011).
94. Vázquez-Suñé, E. *et al.* A geological model for the management of subsurface data in the urban environment of Barcelona and surrounding area. *Solid Earth* **7**, 1317–1329 (2016).
95. Criollo, R. A. An approach for hydrogeological data management, integration and analysis. *Dr. Thesis* (2019).
96. Alcaraz, M. GIS platform for management of shallow geothermal resources. *Doctoral Thesis* (Universitat Politècnica de Catalunya, 2016).
97. Velasco, D. V. GIS-based hydrogeological platform for sedimentary media. *Doctoral Thesis* (Universitat Politècnica de Catalunya, 2013).
98. Microsoft. Database Software and Applications | Microsoft Access. (2019). Available at: <https://www.microsoft.com/en-ww/microsoft-365/access>. (Accessed: 14th September 2021)
99. Esri. ArcGIS. (2021). Available at: <https://www.arcgis.com/index.html>. (Accessed: 14th September 2021)
100. QGIS Development Team. QGIS Geographic Information System. (2021). Available at: <https://qgis.org/en/site/index.html>. (Accessed: 14th September 2021)

101. Velasco, V. *et al.* GIS-based hydrogeochemical analysis tools (QUIMET). *Comput. Geosci.* **70**, 164–180 (2014).
102. Alcaraz, M., García-Gil, A., Vázquez-Suñé, E. & Velasco, V. Use rights markets for shallow geothermal energy management. *Appl. Energy* **172**, 34–46 (2016).
103. Criollo, R. *et al.* An integrated GIS-based tool for aquifer test analysis. *Environ. Earth Sci.* **2016** 755 **75**, 1–11 (2016).
104. OneGeology. (2012). Available at: <http://www.onegeology.org>. (Accessed: 21st April 2021)
105. OGC Standards | OGC. (2012). Available at: <https://www.ogc.org/docs/is>. (Accessed: 21st April 2021)
106. *OGC WaterML 2.0*. (2012).
107. *D2.9 Draft Guidelines for the use of Observations & Measurements and Sensor Web Enablement-related standards in INSPIRE Annex II and III data specification development*. (2013).
108. Hutton, J. *Theory of the Earth*. (1795).
109. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. & Eng.* **9**, 90–95 (2007).
110. Editing polygons in 3D—ArcMap | Documentation. (2016). Available at: <https://desktop.arcgis.com/en/arcmap/latest/extensions/3d-analyst/editing-polygons-in-3d.htm>. (Accessed: 21st April 2021)
111. Aléon, J. *et al.* Determination of the initial hydrogen isotopic composition of the solar system. *Nat. Astron.* **2022** 1–6 (2022). doi:10.1038/s41550-021-01595-7
112. Custodio, E. & Llamas, M. R. *Hidrología Subterránea*. (Omega, 1983).
113. Custodio, E. & Jódar Bermúdez, J. *Recarga natural a los acuíferos, metodología y soporte de la isotopía del agua*. (2019).
114. Gonfiantini, R., Roche, M. A., Olivry, J. C., Fontes, J. C. & Zuppi, G. M. The altitude effect on the isotopic composition of tropical rains. *Chem. Geol.* **181**, 147–167 (2001).
115. Merlivat, L. & Jouzel, J. Global climatic interpretation of the deuterium-oxygen 16 relationship for precipitation. *J. Geophys. Res.* **84**, 5029–5033 (1979).
116. Araguás-Araguás, L., Froehlich, K. & Rozanski, K. Deuterium and oxygen-18 isotope composition of precipitation and atmospheric moisture. in *Hydrological Processes* **14**, 1341–1355 (2000).

117. Jasechko, S. Global Isotope Hydrogeology—Review. *Rev. Geophys.* **57**, 835–965 (2019).
118. Hurley, J. V. & Galewsky, J. A last-saturation diagnosis of subtropical water vapor response to global warming. *Geophys. Res. Lett.* **37**, (2010).
119. Galewsky, J. & Samuels-Crow, K. Summertime moisture transport to the southern South American Altiplano: Constraints from in situ measurements of water vapor isotopic composition. *J. Clim.* **28**, 2635–2649 (2015).
120. Risi, C., Bony, S. & Vimeux, F. Influence of convective processes on the isotopic composition ( $\delta^{18}\text{O}$  and  $\delta\text{D}$ ) of precipitation and water vapor in the tropics: 2. Physical interpretation of the amount effect. *J. Geophys. Res. Atmos.* **113**, (2008).
121. Tharammal, T., Bala, G. & Noone, D. Impact of deep convection on the isotopic amount effect in tropical precipitation. *J. Geophys. Res.* **122**, 1505–1523 (2017).
122. Vimeux, F., Tremoy, G., Risi, C. & Gallaire, R. A strong control of the South American SeeSaw on the intra-seasonal variability of the isotopic composition of precipitation in the Bolivian Andes. *Earth and Planetary Science Letters* **307**, 47–58 (2011).
123. Bailey, A., Posmentier, E. & Feng, X. Patterns of Evaporation and Precipitation Drive Global Isotopic Changes in Atmospheric Moisture. *Geophys. Res. Lett.* **45**, 7093–7101 (2018).
124. Craig, H. Isotopic variations in meteoric waters. *Science (80-. )*. **133**, 1702–1703 (1961).
125. Feng, X., Faiia, A. M. & Posmentier, E. S. Seasonality of isotopes in precipitation: A global perspective. *J. Geophys. Res. Atmos.* **114**, (2009).
126. Gat, J. R. Atmospheric water balance—the isotopic perspective. *Hydrol. Process.* **14**, 1357–1369 (2000).
127. Gat, J. R. & Matsui, E. Atmospheric water balance in the Amazon Basin: an isotopic evapotranspiration model. *J. Geophys. Res.* **96**, 13179–13188 (1991).
128. Salati, E., Dall’Olio, A., Matsui, E. & Gat, J. R. Recycling of water in the Amazon Basin: An isotopic study. *Water Resour. Res.* **15**, 1250–1258 (1979).
129. Thomas, J. M. & Rose, T. P. *Environmental isotopes in hydrogeology*. *Environmental Geology* **43**, (2003).
130. Cook, P. G. & Herczeg, A. L. *Environmental Tracers in Subsurface Hydrology*. *Environmental Tracers in Subsurface Hydrology* (Springer US, 2000). doi:10.1007/978-1-4615-4557-6

131. Coplen, T. *Stable Isotope Hydrology: Deuterium and Oxygen-18 in the Water Cycle*. *Eos, Transactions American Geophysical Union* **63**, (International Atomic Energy Agency, 1982).
132. Kendall, C. & McDonnell, J. J. *Isotope tracers in catchment hydrology*. *Isotope tracers in catchment hydrology* (Elsevier, 1998). doi:10.1029/99eo00193
133. Mook, W. G. *Environmental isotopes in the hydrological cycle Volume I.pdf. Technical documents in hydrology* **1**, (2000).
134. Xi, X. A Review of Water Isotopes in Atmospheric General Circulation Models: Recent Advances and Future Prospects. *Int. J. Atmos. Sci.* **2014**, 1–16 (2014).
135. Wong, T. E., Nusbaumer, J. & Noone, D. C. Evaluation of modeled land-atmosphere exchanges with a comprehensive water isotope fractionation scheme in version 4 of the Community Land Model. *J. Adv. Model. Earth Syst.* **9**, 978–1001 (2017).
136. Nusbaumer, J., Wong, T. E., Bardeen, C. & Noone, D. Evaluating hydrological processes in the Community Atmosphere Model Version 5 (CAM5) using stable isotope ratios of water. *J. Adv. Model. Earth Syst.* **9**, 949–977 (2017).
137. Neale, R. B. *et al.* Description of the NCAR Community Atmosphere Model ( CAM 5 . 0 ). *Ncar/Tn-464+Str* 214 (2004).
138. Steiger, N. J., Steig, E. J., Dee, S. G., Roe, G. H. & Hakim, G. J. Climate reconstruction using data assimilation of water isotope ratios from ice cores. *J. Geophys. Res.* **122**, 1545–1568 (2017).
139. Werner, M., Langebroek, P. M., Carlsen, T., Herold, M. & Lohmann, G. Stable water isotopes in the ECHAM5 general circulation model: Toward high-resolution isotope modeling on a global scale. *J. Geophys. Res. Atmos.* **116**, 15109 (2011).
140. Kurita, N. *et al.* Intraseasonal isotopic variation associated with the Madden-Julian Oscillation. *J. Geophys. Res. Atmos.* **116**, 24101 (2011).
141. Risi, C., Bony, S., Vimeux, F. & Jouzel, J. Water-stable isotopes in the LMDZ4 general circulation model: Model evaluation for present-day and past climates and applications to climatic interpretations of tropical isotopic records. *J. Geophys. Res. Atmos.* **115**, 12118 (2010).
142. Tsuchihara, T., Shirahata, K., Ishida, S. & Yoshimoto, S. Application of a self-organizing map of isotopic and chemical data for the identification of groundwater recharge sources in Nasunogahara alluvial fan, Japan. *Water (Switzerland)* **12**, 278 (2020).

143. Fiorella, R. P. *et al.* Spatiotemporal variability of modern precipitation  $\delta^{18}\text{O}$  in the central Andes and implications for paleoclimate and paleoaltimetry estimates. *J. Geophys. Res.* **120**, 4630–4656 (2015).
144. Garcia, M., Villalba, F., Araguas Araguas, L. & Rozanski, K. The role of atmospheric circulation patterns in controlling the regional distribution of stable isotope contents in precipitation: Preliminary results from two transects in the Ecuadorian Andes. in *Isotope techniques in the study of environmental change. Proceedings of a symposium, Vienna, April 1997.* 127–140 (1998).
145. Guo, X., Tian, L., Wen, R., Yu, W. & Qu, D. Controls of precipitation  $\delta^{18}\text{O}$  on the northwestern Tibetan Plateau: A case study at Ngari station. *Atmos. Res.* **189**, 141–151 (2017).
146. Li, L. & Garzione, C. N. Spatial distribution and controlling factors of stable isotopes in meteoric waters on the Tibetan Plateau: Implications for paleoelevation reconstruction. *Earth Planet. Sci. Lett.* **460**, 302–314 (2017).
147. Nguyen, L. D., Heidbüchel, I., Meyer, H., Merz, B. & Apel, H. What controls the stable isotope composition of precipitation in the Asian monsoon region? *Hydrol. Earth Syst. Sci. Discuss.* 1–33 (2017). doi:10.5194/hess-2017-164
148. Ren, W., Yao, T. & Xie, S. Key drivers controlling the stable isotopes in precipitation on the leeward side of the central Himalayas. *Atmos. Res.* **189**, 134–140 (2017).
149. Rozanski, K., Sonntag, C. & Munnich, K. O. Factors controlling stable isotope composition of European precipitation. *Tellus* **34**, 142–150 (1982).
150. Liebmann, B. Description of a complete (interpolated) outgoing longwave radiation dataset. *Bull. Amer. Meteor. Soc.* **77**, 1275–1277 (1996).
151. Morales, M. S., Christie, D. A., Neukom, R., Rojas, F. & Villalba, R. Variabilidad hidroclimática en el sur del Altiplano: pasado, presente y futuro. *La Puna argentina Nat. y Cult.* **24**, 75–91 (2018).
152. Risi, C. *et al.* What controls the isotopic composition of the African monsoon precipitation? Insights from event-based precipitation collected during the 2006 AMMA field campaign. *Geophys. Res. Lett.* **35**, 1–6 (2008).
153. Vuille, M. *et al.* Climate change and tropical Andean glaciers: Past, present and future. *Earth-Science Rev.* **89**, 79–96 (2008).
154. Nguyen, L. D., Heidbüchel, I., Meyer, H., Merz, B. & Apel, H. What controls the stable isotope

- composition of precipitation in the Asian monsoon region? *Hydrol. Earth Syst. Sci. Discuss.* 1–33 (2017). doi:10.5194/HESS-2017-164
155. Stein, A. F. *et al.* NOAA’s HYSPLIT Atmospheric Transport and Dispersion Modeling System. *Bull. Am. Meteorol. Soc.* **96**, 2059–2077 (2015).
  156. Muñoz-Sabater, J. *et al.* ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data* **13**, 4349–4383 (2021).
  157. Schmidt, G. A. *et al.* Present-Day Atmospheric Simulations Using GISS ModelE: Comparison to In Situ, Satellite, and Reanalysis Data. *J. Clim.* **19**, 153–192 (2006).
  158. Yoshimura, K., Kanamitsu, M., Noone, D. & Oki, T. Historical isotope simulation using Reanalysis atmospheric data. *J. Geophys. Res. Atmos.* **113**, 19108 (2008).
  159. Koh, K., Kim, S. J. & Boyd, S. A method for large-scale  $\ell_1$ -regularized logistic regression. in *Proceedings of the National Conference on Artificial Intelligence* **1**, 565–571 (2007).
  160. Tipping, M. E. Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **1**, 211–244 (2001).
  161. Efron, B. *et al.* Least angle regression. <https://doi.org/10.1214/009053604000000067> **32**, 407–499 (2004).
  162. MacKay, D. J. C. Bayesian nonlinear modeling for the prediction competition. in *ASHRAE Transactions* **100**, 1053–1062 (1994).
  163. Mallat, S. G. & Zhang, Z. Matching Pursuits With Time-Frequency Dictionaries. *IEEE Trans. Signal Process.* **41**, 3397–3415 (1993).
  164. Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. large margin Classif.* **10**, 61–74 (1999).
  165. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* 2006 631 **63**, 3–42 (2006).
  166. Hinton, G. E. Connectionist Learning Procedures. (1989).
  167. Claesen, M. & De Moor, B. Hyperparameter Search in Machine Learning. (2015).
  168. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. in *Proceedings of the IEEE International Conference on*

- Computer Vision* **2015 Inter**, 1026–1034 (2015).
169. Cranganu, C. & Breaban, M. Using support vector regression to estimate sonic log distributions: A case study from the Anadarko Basin, Oklahoma. *J. Pet. Sci. Eng.* **103**, 1–13 (2013).
  170. Chang, C. C. & Lin, C. J. LIBSVM: A Library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**, (2011).
  171. Wu, T.-F., Lin, C.-J. & Weng, R. C. Probability Estimates for Multi-class Classification by Pairwise Coupling. *J. Mach. Learn. Res.* **5**, 975–1005 (2004).
  172. Ploton, P. *et al.* Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nat. Commun.* **11**, 1–11 (2020).
  173. Kraskov, A., Stögbauer, H. & Grassberger, P. Estimating mutual information. *Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top.* **69**, 16 (2004).
  174. Li, W. Mutual information functions versus correlation functions. *J. Stat. Phys.* **60**, 823–837 (1990).
  175. Smith, R. A mutual information approach to calculating nonlinearity. *Stat* **4**, 291–303 (2015).
  176. The pandas development team. pandas-dev/pandas: Pandas. (2020). doi:10.5281/zenodo.3509134
  177. Jordahl, K. *et al.* geopandas/geopandas: v0.10.2. (2021). doi:10.5281/ZENODO.5573592
  178. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
  179. Haentjens, N. pylr2 · PyPI. (2018). Available at: <https://pypi.org/project/pylr2/>. (Accessed: 3rd March 2022)
  180. Seabold, S. & Perktold, J. statsmodels: Econometric and statistical modeling with python. in *9th Python in Science Conference* (2010).
  181. González Rouco, J., Jiménez, J., Quesada, V. & Valero Rodríguez, F. Quality control and homogeneity of precipitation data in the southwest of Europe. *J. Clim.* **14**, 964–978 (2001).
  182. Peterson, T. C., Vose, R., Schmoyer, R. & Razuvaëv, V. Global historical climatology network (GHCN) quality control of monthly temperature data. *Int. J. Climatol.* **18**, 1169–1179 (1998).
  183. Somaya, H. & Tomader, M. Tuning the hyperparameters for supervised machine learning classification, to optimize detection of IoT Botnet. 1–6 (2022). doi:10.1109/ISIVC54825.2022.9800742

184. Gillies, S. & others. Shapely: manipulation and analysis of geometric objects. (2007).
185. Bokeh Development Team. Bokeh: Python library for interactive visualization. (2018).
186. Friedman, J., Bohonak, A. J. & Levine, R. A. When are two pieces better than one: fitting and testing OLS and RMA regressions. *undefined* **24**, 306–316 (2013).
187. Chen, F. *et al.* Local Meteoric Water Lines in a Semi-Arid Setting of Northwest China Using Multiple Methods. *Water* **2021**, Vol. 13, Page 2380 **13**, 2380 (2021).
188. Crawford, J., Hughes, C. E. & Lykoudis, S. Alternative least squares methods for determining the meteoric water line, demonstrated using GNIP data. *J. Hydrol.* **519**, 2331–2340 (2014).
189. McKerns, M., Strand, L., Sullivan, T., Fang, A. & Aivazis, M. Building a Framework for Predictive Science. in *Proceedings of the 10th Python in Science Conference* 76–86 (2011). doi:10.25080/majora-ebaa42b7-00d
190. Amphos21. *Estudio de modelos hidrogeológicos conceptuales integrados, para los salares de Atacama, Maricunga y Pedernales. Etapa III. Informe Final. Modelo Hidrogeológico Consolidad Cuenca Salar de Atacama.* (2018).
191. DGA. *Análisis de la Oferta Hídrica del Salar de Atacama. Sdt N° 339* (2013).
192. DGA. *Evaporación desde salares: Metodología para evaluar los recursos hídricos renovables. Aplicación en las regiones I y II. Revista de la Sociedad Chilena de Ingeniería Hidráulica* **1**, (1986).
193. Hess, R. A. *Simplified approach for modelling pilot pursuit control behaviour in multi-loop flight control tasks. Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering* **220**, (2006).
194. Marazuela, M. A., Vázquez-Suñé, E., Ayora, C. & García-Gil, A. Towards more sustainable brine extraction in salt flats: Learning from the Salar de Atacama. *Sci. Total Environ.* **703**, 135605 (2020).
195. Marazuela, M. A., Vázquez-Suñé, E., Ayora, C., García-Gil, A. & Palma, T. The effect of brine pumping on the natural hydrodynamics of the Salar de Atacama: The damping capacity of salt flats. *Sci. Total Environ.* **654**, 1118–1131 (2019).
196. Marazuela, M. A. *et al.* 3D mapping, hydrodynamics and modelling of the freshwater-brine mixing zone in salt flats similar to the Salar de Atacama (Chile). *J. Hydrol.* **561**, 223–235 (2018).
197. IDAEA-CSIC. *Cuarta actualización del modelo hidrogeológico del Salar de Atacama. SNIFA*<https://snifa.sma.gob.cl>. (2017).



198. Valdivielso, S., Vázquez-Suñé, E., Herrera, C. & Custodio, E. Characterization of precipitation and recharge in the peripheral aquifer of the Salar de Atacama. *Sci. Total Environ.* **806**, 150271 (2022).
199. DGA. Servicios hidrometeorológicos. [www.dga.cl/servicioshidrometeorologicos/Paginas/default.aspx](http://www.dga.cl/servicioshidrometeorologicos/Paginas/default.aspx) (2020).
200. Centre for Climate and Resilience Research. Datos de precipitación, Datos de temperaturas. [www.cr2.cl/datos-de-precipitacion/](http://www.cr2.cl/datos-de-precipitacion/) (2018).
201. Cortecchi, G. *et al.* New chemical and original isotopic data on waters from El Tatio geothermal field, northern Chile. *Geochem. J.* **39**, 547–571 (2005).
202. CRICYT. *Segundo informe de avance sobre estudios e investigaciones que intentan explicar el estado actual de ejemplares de Algarrobo, en una población ubicada en las proximidades del pozo CAMAR 2 de SQM, en el Salar de Atacama, Chile.* (2017). doi:10.1079/BJN20041276
203. Herrera, C. *et al.* Groundwater flow in a closed basin with a saline shallow lake in a volcanic area: Laguna Tuyajto, northern Chilean Altiplano of the Andes. *Sci. Total Environ.* **541**, 303–318 (2016).
204. Lagos Durán, L. V. Hidrogeoquímica de fuentes termales en ambientes salinos relacionados con salares en los Andes del norte de Chile. *MSc Thesis* (Universidad de Chile. Tesis for degree of Master of Sciences mention in geology, 2016).
205. Moran, B. J., Boutt, D. F. & Munk, L. A. Stable and Radioisotope Systematics Reveal Fossil Water as Fundamental Characteristic of Arid Orogenic-Scale Groundwater Systems. *Water Resour. Res.* **55**, 11295–11315 (2019).
206. Valdivielso, S., Hassanzadeh, A., Vázquez-Suñé, E., Custodio, E. & Criollo, R. Spatial Distribution of Meteorological Factors Controlling Stable Isotopes in Precipitation in Northern Chile. *J. Hydrol.* **605**, (2022).
207. DICTUC. *Levantamiento hidrogeológico para el desarrollo de nuevas fuentes de agua en áreas prioritarias de la zona norte de Chile, regiones XV, I, II y III. Etapa 2. Informe Final parte IX. Sistema Hidrogeoquímica e isotopía regional del Altiplano de Chile. Sistem. Parte IX*, (2009).
208. Houston, J. & Hartley, A. J. The central andean west-slope rainshadow and its potential contribution to the origin of hyper-aridity in the Atacama Desert. *Int. J. Climatol.* **23**, 1453–1464 (2003).
209. Marazuela, M. A., Vázquez-Suñé, E., Ayora, C., García-Gil, A. & Palma, T. Hydrodynamics of salt flat basins: The Salar de Atacama example. *Sci. Total Environ.* **651**, 668–683 (2019).

210. Vazquez, Enric & Ayora, C. *Cuarta actualización del modelo hidrogeológico del Salar de Atacama*. (2017).
211. Kampf, S. K., Tyler, S. W., Ortiz, C. A., Muñoz, J. F. & Adkins, P. L. Evaporation and land surface energy budget at the Salar de Atacama, Northern Chile. *J. Hydrol.* **310**, 236–252 (2005).
212. MOP-DGA. *Balance Hídrico de América del Sur*. (1988).
213. Valdivielso, S., Vázquez-Suñé, E. & Custodio, E. Origin and variability of oxygen and hydrogen isotopic composition of precipitation in the Central Andes: A review. *J. Hydrol.* **587**, 124899 (2020).
214. Valdivielso, S., Vázquez-Suñé, E. & Custodio, E. Environmental isotope concepts of precipitation and surface water and groundwater in the central andes: A review. *Bol. Geol. y Min.* **132**, 147–156 (2021).
215. SGA. *Estudio Hidrogeológico y Modelo Numérico sector sur del Salar de Atacama*. (2015).
216. Salas, J., Moreno, R., Moreno, R. & Bruno, J. *Interpretación y contexto hidrogeológico de los puntos de control del Plan de Contingencia del Sistema Soncor . Análisis de su representatividad*. (2010).
217. Valdivielso, S., Vázquez-Suñé, E., Herrera, C. & Custodio, E. Characterization of precipitation and recharge in the peripheral aquifer of the Salar de Atacama. *Sci. Total Environ.* **806**, 150271 (2022).
218. Geol, X. I. I. C., Santiago, C., Geol, C., Ambiente, M. & Cient, S. Estudio de la relación isotópica  $\delta^{18}O / \delta^2H$  de los manantiales en el sector de las nacientes del Loa , Región de Antofagasta. *XII Congr. Geológico Chil.* 16–19 (2009).
219. Villablanca, D. Estudio de la relación isotópica  $\delta^{18}O/\delta^2H$  de los manantiales en el sector de las nacientes del Loa, Región de Antofagasta. *XII Congr. Geológico Chil.* 22–26 (2009).
220. Valdivielso, S., Hassanzadeh, A., Vázquez-Suñé, E., Custodio, E. & Criollo, R. Spatial distribution of meteorological factors controlling stable isotopes in precipitation in Northern Chile. *J. Hydrol.* **605**, 127380 (2022).
221. Fritz, P., Suzuki, O., Silva, C. & Salati, E. Isotope hydrology of groundwaters in the Pampa del Tamarugal, Chile. *J. Hydrol.* **53**, 161–184 (1981).
222. Aravena, R. *et al.* Isotopic composition and origin of the precipitation in Northern Chile. *Appl. Geochemistry* **14**, 411–422 (1999).
223. Chaffaut, I., Coudrain-Ribstein, A., Michelot, J. L. & Pouyaud, B. Précipitation d'altitude du nord-

- Chili, origine des sources de vapeur et données isotopiques. *Bull. l'Institut français d'études Andin.* **27**, 367–384 (1998).
224. Chaffaut, I. Precipitations d'altitude, eaux souterraines et changements climatiques de l'Altiplano nord-Chilien. (PhD Thesis of Université Paris Sud U.F.R. Scientifique D'Orsay., 1998).
225. Boschetti, T., Cifuentes, J., Iacumin, P. & Selmo, E. Local meteoric water line of northern Chile (18° S-30° S): An application of error-in-variables regression to the oxygen and hydrogen stable isotope ratio of precipitation. *Water (Switzerland)* **11**, (2019).
226. Verma, R. K., Verma, S., Mishra, S. K. & Pandey, A. SCS-CN-Based Improved Models for Direct Surface Runoff Estimation from Large Rainfall Events. *Water Resour. Manag.* **35**, 2149–2175 (2021).
227. Sophocleous, M. Interactions between groundwater and surface water: The state of the science. *Hydrogeol. J.* **10**, 52–67 (2002).
228. Vinet, L. & Zhedanov, A. A 'missing' family of classical orthogonal polynomials. *J. Phys. A Math. Theor.* **44**, 343–354 (2011).
229. Hauwert, N. M., Sharp, J. M., Hauwert, N. M. & Sharp, J. M. Measuring Autogenic Recharge over a Karst Aquifer Utilizing Eddy Covariance Evapotranspiration. *J. Water Resour. Prot.* **6**, 869–879 (2014).
230. Healy, R. W. & Cook, P. G. Using groundwater levels to estimate recharge. *Hydrogeol. J.* **10**, 91–109 (2002).
231. Bellot, J. & Chirino, E. Hydrobal: An eco-hydrological modelling approach for assessing water balances in different vegetation types in semi-arid areas. *Ecol. Modell.* **266**, 30–41 (2013).
232. U.S. Geological Survey. Soil-Water-Balance (SWB). (2019).
233. US Army Corps of Engineers. HEC-HMS. (2022).
234. Mckeever, V., Owen, W., Rallison, R. & Engineers, H. *NATIONAL ENGINEERING HANDBOOK SECTION 4.* (1965).
235. Ponce, V. M. & Hawkins, R. H. Runoff Curve Number: Has It Reached Maturity? *J. Hydrol. Eng.* **1**, 11–19 (1996).
236. Hawkins, H., Ward, T. J., Woodward, D. E., Van Mullem, J. A. & McCuen, R. H. Review of Curve Number Hydrology: State of the Practice by R. H. Hawkins, T. J. Ward, D. E. Woodward, and J. A.

- Van Mullem. *J. Hydrol. Eng.* **14**, 1046–1047 (2009).
237. Aragaw, H. M. & Mishra, S. K. Clarification of issues and long-duration hydrologic simulation SCS-CN-based proxy modelling. *Acta Geophys.* **70**, 729–756 (2022).
238. Department of Agriculture, U. S., Natural Resources Conservation Service & Conservation Engineering Division. *Urban Hydrology for Small Watersheds TR-55*. (1986).
239. Ajmal, M., Waseem, M., Ahn, J.-H. & Kim, T.-W. Runoff Estimation Using the NRCS Slope-Adjusted Curve Number in Mountainous Watersheds. *J. Irrig. Drain. Eng.* **142**, (2016).
240. Hawkins, R. H., Theurer, F. D. & Rezaeianzadeh, M. Understanding the Basis of the Curve Number Method for Watershed Models and TMDLs. *J. Hydrol. Eng.* **24**, (2019).
241. Ajmal, M., Waseem, M., Kim, D. & Kim, T. W. A Pragmatic Slope-Adjusted Curve Number Model to Reduce Uncertainty in Predicting Flood Runoff from Steep Watersheds. *Water* **2020**, Vol. 12, Page 1469 **12**, 1469 (2020).
242. Chakraborty, S., Pandey, R. P., Mishra, S. K. & Chaube, U. C. Relation Between Runoff Curve Number and Irrigation Water Requirement. *Agric. Res.* **4**, 378–387 (2015).
243. Chin, D. A. On relationship between curve numbers and phi indices. *Water Sci. Eng.* **11**, 187–195 (2018).
244. Hooshyar, M. & Wang, D. An analytical solution of Richards' equation providing the physical basis of SCS curve number method and its proportionality relationship. *Water Resour. Res.* **52**, 6611–6620 (2016).
245. Satheeshkumar, S., Venkateswaran, S & Kannan, R. Rainfall–runoff estimation using SCS–CN and GIS approach in the Pappiredipatti watershed of the Vaniyar sub basin, South India. *Model. Earth Syst. Environ.* **2017 31 3**, 1–8 (2017).
246. Shi, W., Huang, M., Gongadze, K. & Wu, L. A Modified SCS-CN Method Incorporating Storm Duration and Antecedent Soil Moisture Estimation for Runoff Prediction. *Water Resour. Manag.* **31**, 1713–1727 (2017).
247. Verma, S., Mishra, S. K., Singh, A., Singh, P. K. & Verma, R. K. An enhanced SMA based SCS-CN inspired model for watershed runoff prediction. *Environ. Earth Sci.* **76**, (2017).
248. Yuan, Y., Mitchell, J. K., Hirsch, M. C. & Cooke, R. A. C. Modified SCS curve number method for predicting subsurface drainage flow. *Trans. Am. Soc. Agric. Eng.* **44**, 1673–1682 (2001).

249. Zhou, S.-M., Warrington, D. N., Lei, T.-W., Lei, Q.-X. & Zhang, M.-L. Modified CN Method for Small Watershed Infiltration Simulation. *J. Hydrol. Eng.* **20**, 04014095 (2014).
250. Mishra, S. K., Jain, M. K., Suresh Babu, P., Venugopal, K. & Kaliappan, S. Comparison of AMC-dependent CN-conversion formulae. *Water Resour. Manag.* **22**, 1409–1420 (2008).
251. Shi, W. & Wang, N. An Improved SCS-CN Method Incorporating Slope, Soil Moisture, and Storm Duration Factors for Runoff Prediction. *Water* 2020, Vol. 12, Page 1335 **12**, 1335 (2020).
252. Pathiraja, S., Westra, S. & Sharma, A. Why continuous simulation? the role of antecedent moisture in design flood estimation. *Water Resour. Res.* **48**, (2012).
253. Woldemeskel, F. & Sharma, A. Should flood regimes change in a warming climate? The role of antecedent moisture conditions. *Geophys. Res. Lett.* **43**, 7556–7563 (2016).
254. Ministry of Agriculture - British Columbia. *SOIL WATER STORAGE CAPACITY AND AVAILABLE SOIL MOISTURE*. (2015).
255. Blaney, H. F. & Criddle, W. D. *Determining Water Requirements in Irrigated Area from Climatological Irrigation Data*. US Dep. of Agr. (1950).
256. Hamon, W. R. & AM.ASCE. Estimating Potential Evapotranspiration. *Trans. Am. Soc. Civ. Eng.* **128**, 324–338 (1963).
257. Linacre, E. T. A simple formula for estimating evaporation rates in various climates, using temperature data alone. *Agric. Meteorol.* **18**, 409–424 (1977).
258. Romanenko, V. A. Computation of the Autumn Soil Moisture Using a Universal Relationship for a Large Area. in *Ukrainian Hydrometeorological Research Institute* (1961).
259. Abtew, W. EVAPOTRANSPIRATION MEASUREMENTS AND MODELING FOR THREE WETLAND SYSTEMS IN SOUTH FLORIDA1. *JAWRA J. Am. Water Resour. Assoc.* **32**, 465–473 (1996).
260. FAO. *Crop water requirements*. (1992).
261. Hargreaves, G. H. & Samani, Z. A. Estimating potential evapotranspiration. *J. Irrig. Drain. Div.* **108**, (1982).
262. Jensen, M. E. & Haise, H. R. Estimating Evapotranspiration from Solar Radiation. *J. Irrig. Drain. Div.* **89**, 15–41 (1963).

263. Xu, C.-Y. & Singh, V. P. Cross Comparison of Empirical Equations for Calculating Potential Evapotranspiration with Data from Switzerland. *Water Resour. Manag.* **16**, 197–219 (2002).
264. McGuinness, J. L. & Borone, E. F. A Comparison of Lysimeter-Derived Potential Evapotranspiration With Computed Values. (1972). doi:10.22004/AG.ECON.171893
265. Flores, N. *et al.* Comparison of Three Daily Rainfall-Runoff Hydrological Models Using Four Evapotranspiration Models in Four Small Forested Watersheds with Different Land Cover in South-Central Chile. *Water 2021, Vol. 13, Page 3191* **13**, 3191 (2021).
266. Trajkovic, S. & Kolakovic, S. Evaluation of reference evapotranspiration equations under humid conditions. *Water Resour. Manag.* **23**, 3057–3067 (2009).
267. Wright, J. L. New Evapotranspiration Crop Coefficients. *J. Irrig. Drain. Div.* **108**, 57–74 (1982).
268. PENMAN, H. L. Natural evaporation from open water, bare soil and grass. *Proc. R. Soc. Lond. A. Math. Phys. Sci.* **193**, 120–145 (1948).
269. Allen, R. G., Pereira, L. S., Raes, D. & Smith, M. Crop evapotranspiration-guidelines for computing crop water requirements-FAO irrigation and drainage paper 56. *FAO, Rome* **300**, (1998).
270. Priestley, C. H. B. & Taylor, R. J. On the assessment of surface heat flux and evaporation using large-scale parameters. *Mon. Weather Rev.* **100**, (1972).
271. Thom, A. S. & Oliver, H. R. On Penman's equation for estimating regional evaporation. *Q. J. R. Meteorol. Soc.* **103**, 345–357 (1977).
272. Tubau, I., Vázquez-Suñé, E., Carrera, J., Valhondo, C. & Criollo, R. Quantification of groundwater recharge in urban environments. *Sci. Total Environ.* **592**, 391–402 (2017).
273. Unidata. NetCDF. (2022). doi:10.5065/D6H70CW6
274. Vremec, M., Collenteur, R. A. & Birk, S. Technical note: Improved handling of potential evapotranspiration in hydrological studies with PyEt. *Hydrol. Earth Syst. Sci. Discuss.* 1–23 (2023). doi:10.5194/HESS-2022-417
275. GDAL/OGR contributors. Geospatial Data Abstraction software Library. (2022). doi:10.5281/zenodo.5884351
276. Chaturvedi, R. S. A note on the investigation of ground water resources in western districts of Uttar Pradesh. annual report, U. P. *Irrig. Res. Inst.* **1973**, 86 – 122 (1973).

277. Kumar, C. P. & Seethapathi, P. V. Assessment of natural groundwater recharge in Upper Ganga Canal command area. *J. Appl. Hydrol.* **15**, 13 – 20 (2002).
278. Maxey, G. B. & Eakin, T. E. Ground water in White River Valley, White Pine, Nye, and Lincoln Counties, Nevada. (1949).
279. Rao, K. *Hydrometeorological aspects of estimating groundwater potential. Seminar on Ground Water Potential* (1970).

## Appendices

### A. Additional Figures of Chapter 2

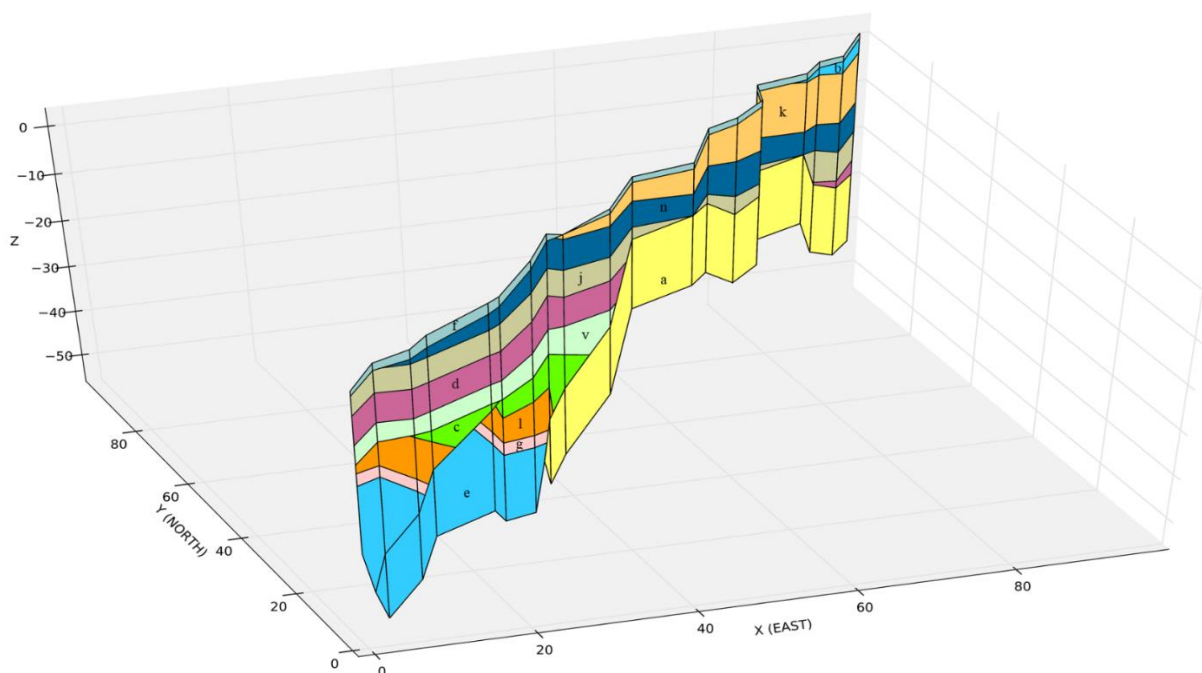


Figure A.34 A 3D cross-section based on synthetic database 2. Note that the geospatial properties of the boreholes change from synthetic dataset 2 to demonstrate the 3D capabilities of Geopropy.

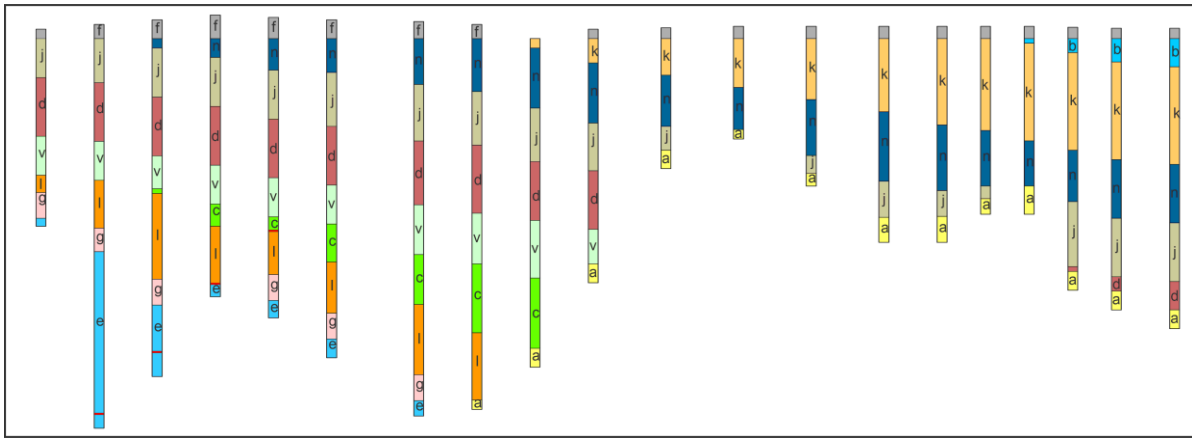


Figure A.35 Raw data of synthetic dataset 2.

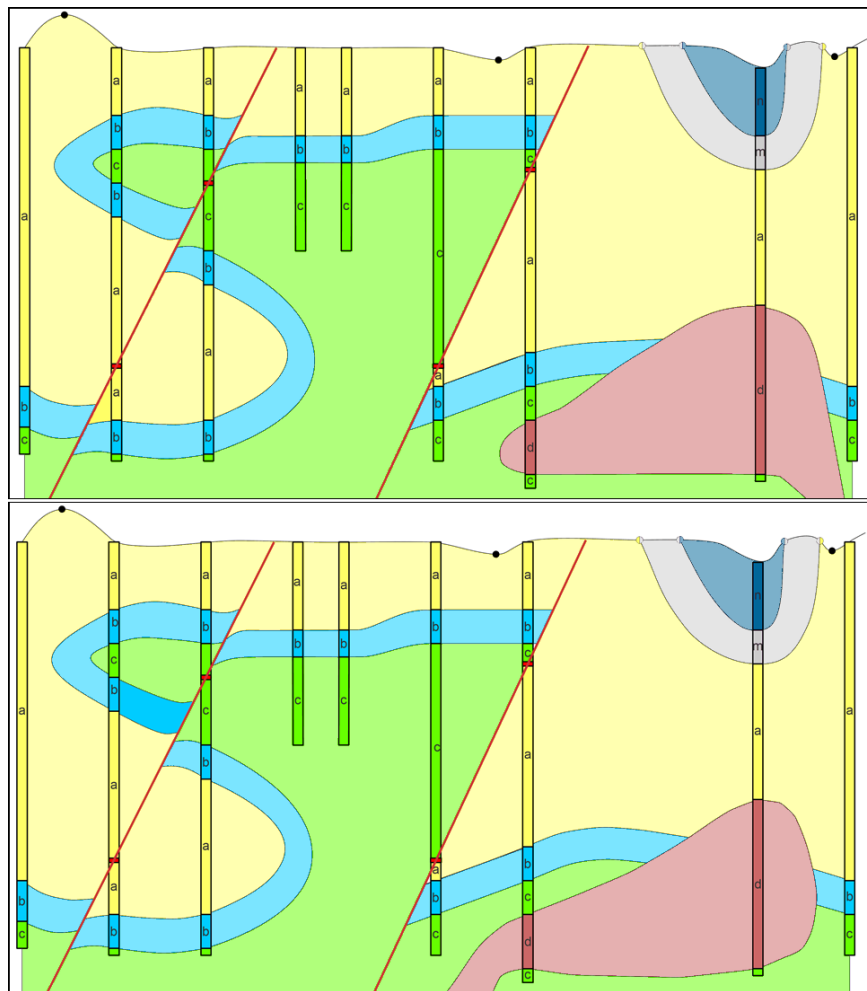


Figure A.36 Cross-sections of synthetic dataset 3, created by the geologist. Based on the available data, the geological units in critical zones *iii* and *iv* could be interpreted in various ways.



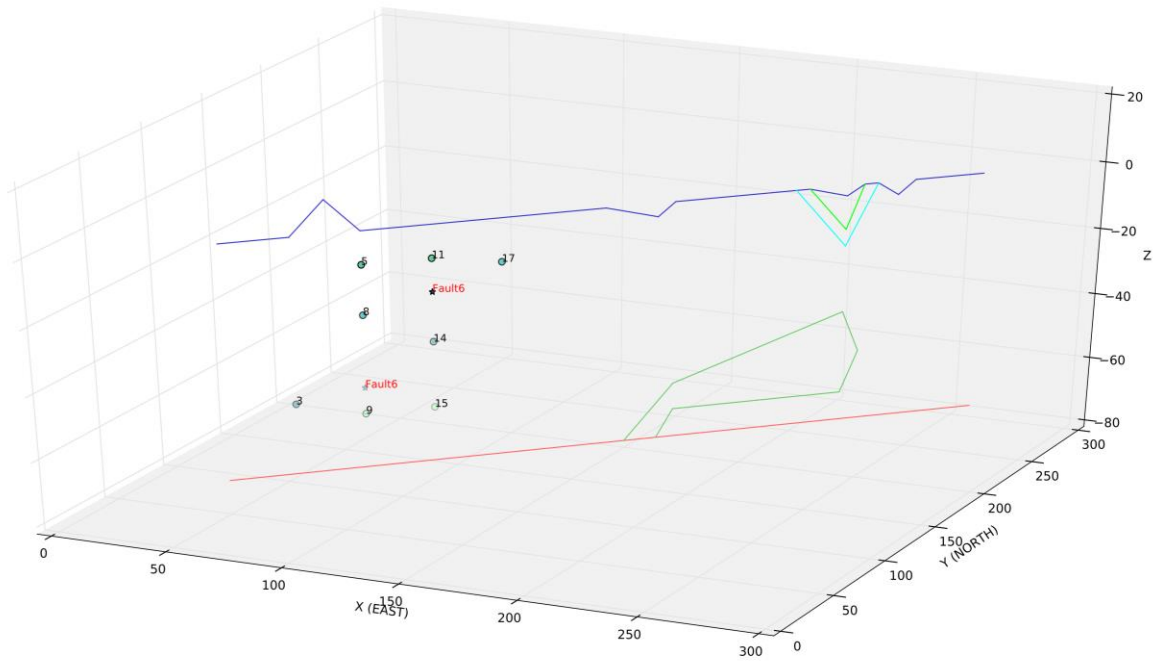


Figure A.37 Screenshot of the semiautomatic-stage visualization of synthetic dataset 3. The interactive visualization tool helps the user identify the faults, contact points and respective point IDs. The provisional scheme of unit contacts that are already created is shown by lines in different colours.

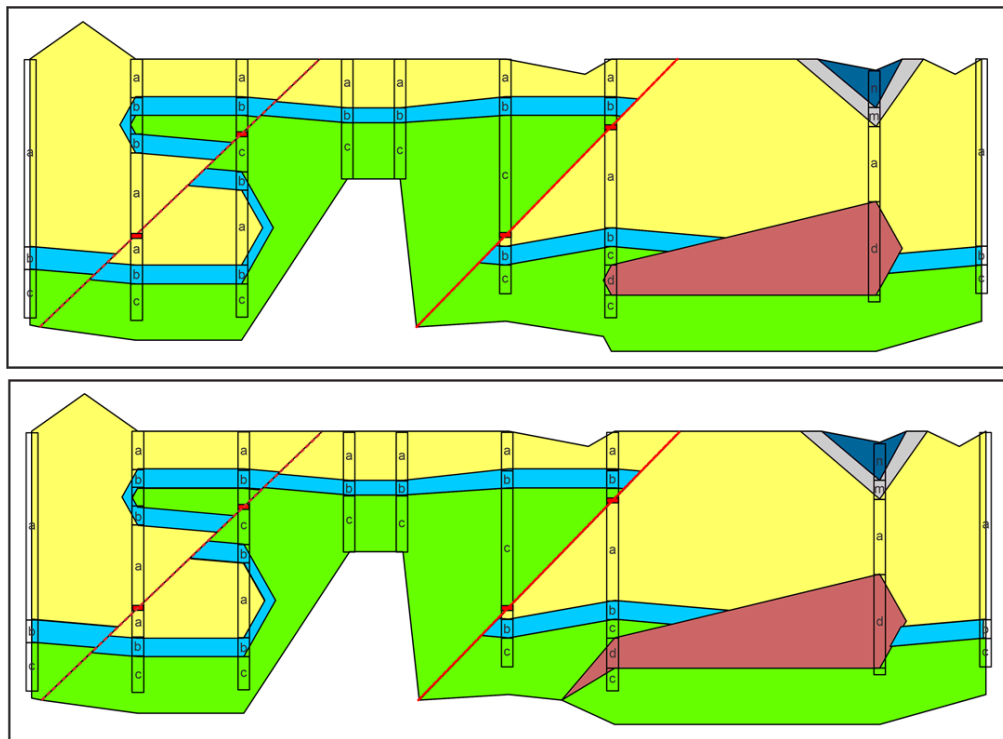


Figure A.38 Cross-sections of synthetic dataset 3, created by Geopropy. The user decided to choose the semiautomatic-stage preferences in critical zones *iii* and *iv*.

## B. Synthetic Datasets Tables of Chapter 2

Table B.1 Borehole data table of the synthetic dataset 1

borehole_id	x	y	elevation
1	0	0	2
2	21	21	1
3	27	27	0
4	40	40	0
5	46	46	0
6	53	53	0
7	62	62	0

Table B.2 Borehole Unit data table of the synthetic dataset 1

borehole_id	top_depth	bottom_depth	units
1	0	2	e
1	2	3	a
1	3	9	b
1	9	11.3	f
2	0	2	a
2	2	2	Fault
2	2	4	a
2	4	9.3	b
2	9.3	11	f
3	0	5	a
3	5	8	b
3	8	9.5	b
3	8	8	Fault

3	9.5	11	f
4	0	6	a
4	6	13.8	b
5	0	5	a
5	5	7	c
5	7	10.3	b
6	0	5	a
6	5	8	c
6	8	12	b
7	0	5	a
7	5	6.2	d
7	6.2	9.1	c
7	9.1	12.1	b
7	12.1	14.1	b

Table B.3 Chronological data table of the synthetic dataset 1

priority_number	bottom_layer	top_layer	type	preferred_angle
1	f	b	conformity	
2	b	c	conformity	
3	c	d	conformity	
4		a	unconformity	
5			fault	
6	a	e	conformity	

Table B.4 Fault data table of the synthetic dataset 1

priority_number	borehole_id	elevation	preferred_angle	type
5	2	2	45	fault
5	3	8	45	fault

Table B.5 Borehole data  
table of the synthetic  
dataset 2

borehole_id	x	y	elevation
111	0	0	1
112	5	5	1.5
113	10	10	2
114	15	15	2.5
115	20	20	2.25
116	25	25	2
117	32.5	32.5	1.8
118	37.5	37.5	1.5
119	42.5	42.5	0
1110	47.5	47.5	1
1111	53.75	53.75	1.15
1112	60	60	1.3
1113	66.25	66.25	1.3
1114	72.5	72.5	1.3
1115	77.5	77.5	1.3
1116	81.25	81.25	1.3
1117	85	85	1.3
1118	88.75	88.75	1.2
1119	92.5	92.5	1.1
1120	97.5	97.5	1.1

Table B.6 Borehole Unit data table  
of the synthetic dataset 2

borehole_id	top_depth	bottom_depth	units
111	0	1	f
111	1	5.16	j
111	5.16	11.4	d
111	11.4	15.56	v
111	15.56	17.4	l
111	17.4	20.15	g
111	20.15	21	e
112	0	1.5	f
112	1.5	6.2	j

112	6.2	12.44	d
112	12.44	16.6	v
112	16.6	21.7	l
112	21.7	24.2	g
112	24.2	41.5	e
112	41.5	41.5	Fault
112	41.5	43	e
113	0	2	f
113	2	3	n
113	3	8.24	j
113	8.24	14.48	d
113	14.48	18	v
113	18	18.5	c
113	18.5	27.64	l
113	27.64	30.39	g
113	30.39	35.39	e
113	35.39	35.39	Fault
113	35.39	38	e
114	0	2.5	f
114	2.5	4.5	n
114	4.5	9.74	j
114	9.74	15.98	d
114	15.98	20.14	v
114	20.14	22.5	c
114	22.5	28.64	l
114	28.64	30	e
114	28.64	28.64	Fault
115	0	2.25	f
115	2.25	5.6	n
115	5.6	10.84	j
115	10.84	17.08	d
115	17.08	21.24	v
115	21.24	22.75	c
115	22.75	22.75	Fault
115	22.75	27.4	l
115	27.4	30.15	g
115	30.15	32	e
116	0	2	f
116	2	5.62	n
116	5.62	11.36	j
116	11.36	17.6	d
116	17.6	21.76	v

116	21.76	25.8	c
116	25.8	31.26	l
116	31.26	34.01	g
116	34.01	36	e
117	0	1.8	f
117	1.8	6.68	n
117	6.68	12.72	j
117	12.72	19.5	d
117	19.5	24.8	v
117	24.8	30.12	c
117	30.12	37.62	l
117	37.62	40.37	g
117	40.37	42	e
118	0	1.5	f
118	1.5	7.14	n
118	7.14	12.86	j
118	12.86	20.1	d
118	20.1	25.5	v
118	25.5	32.84	c
118	32.84	40	l
118	40	41	a
119	0	1	k
119	1	7.4	n
119	7.4	13.12	j
119	13.12	19.36	d
119	19.36	25.52	v
119	25.52	33	c
119	33	35	a
1110	0	1	f
1110	1	3.6	k
1110	3.6	10	n
1110	10	15.08	j
1110	15.08	21.32	d
1110	21.32	25	v
1110	25	27	a
1111	0	1.15	f
1111	1.15	5.05	k
1111	5.05	10.5	n
1111	10.5	13.04	j
1111	13.04	15	a
1112	0	1.3	f
1112	1.3	6.5	k

1112	6.5	11	n
1112	11	12	a
1113	0	1.3	f
1113	1.3	7.8	k
1113	7.8	13.75	n
1113	13.75	15.67	j
1113	15.67	17	a
1114	0	1.3	f
1114	1.3	9.1	k
1114	9.1	16.5	n
1114	16.5	20.34	j
1114	20.34	23	a
1115	0	1.3	f
1115	1.3	10.5	k
1115	10.5	17.5	n
1115	17.5	20.24	j
1115	20.24	23	a
1116	0	1.3	f
1116	1.3	11.1	k
1116	11.1	17	n
1116	17	18.37	j
1116	18.37	20	a
1117	0	1.3	f
1117	1.3	1.8	b
1117	1.8	12.2	k
1117	12.2	17	n
1117	17	20	a
1118	0	1.2	f
1118	1.2	2.7	b
1118	2.7	13.1	k
1118	13.1	18.57	n
1118	18.57	25.5	j
1118	25.5	26.01	d
1118	26.01	28	a
1119	0	1.1	f
1119	1.1	3.6	b
1119	3.6	14	k
1119	14	20.24	n
1119	20.24	26.48	j
1119	26.48	28	d
1119	28	30	a
1120	0	1.1	f

1120	1.1	4.1	b
1120	4.1	14.5	k
1120	14.5	20.74	n
1120	20.74	26.98	j
1120	26.98	30	d
1120	30	32	a

Table B.7 Chronological data table of the synthetic dataset 2

priority_number	bottom_layer	top_layer	type	preferred_angle
2	e	g	conformity	
3	g	l	conformity	
4	l	c	conformity	
5			Fault	
6		v	unconformity	
7	v	d	conformity	
8	d	j	conformity	
9	a		intrusion	
10		n	unconformity	
11	n	k	conformity	
12	k	b	conformity	
13		f	unconformity	

Table B.8 Fault data table of the synthetic dataset 2

priority_number	borehole_id	elevation	preferred_angle	type
5	113	35.39	135	Fault
5	114	28.64	135	Fault
5	112	41.5	135	Fault
5	115	22.75	135	Fault

Table B.9 Borehole data table of the synthetic dataset 3

borehole_id	x	y	elevation
-------------	---	---	-----------



111	70	70	0
112	90	90	0
113	110	110	0
114	130	130	0
115	140	140	0
116	160	160	0
117	180	180	0
118	230	230	-3
119	250	250	0

Table B.10 Borehole Unit data table  
of the synthetic dataset 3

borehole_id	top_depth	bottom_depth	units
111	0	50	a
111	50	56	b
111	56	60	c
112	0	10	a
112	10	15	b
112	15	20	c
112	20	25	b
112	25	47	a
112	47	55	a
112	47	47	Fault
112	55	60	b
112	60	61	c
113	0	10	a
113	10	15	b
113	15	20	c
113	20	30	c
113	20	20	Fault
113	30	35	b
113	35	55	a
113	55	60	b
113	60	61	c
114	0	13	a
114	13	17	b

114	17	61	c
115	0	13	a
115	13	17	b
115	17	55	c
116	0	10	a
116	10	15	b
116	15	47	c
116	47	50	a
116	47	47	Fault
116	50	55	b
116	55	61	c
117	0	10	a
117	10	15	b
117	15	18	c
117	18	45	a
117	18	18	Fault
117	45	50	b
117	50	55	c
117	55	63	d
117	63	65	c
118	0	10	n
118	10	15	m
118	15	35	a
118	35	60	d
118	60	61	c
119	0	50	a
119	50	55	b
119	55	61	c

Table B.11 Chronological data table of the synthetic dataset 3

priority_number	bottom_layer	top_layer	type	preferred_angle
1	c	b	conformity	5
2	b	a	conformity	5
3	d		Intrusion	
4			Fault	

5			Fault	
6	a	m	conformity	
7	m	n	conformity	

Table B.12 Fault data table of the synthetic dataset 3

priority_number	borehole_id	elevation	preferred_angle	type
5	116	45	135	Fault
5	117	18	135	Fault
6	112	47	135	Fault
6	113	20	135	Fault

Table B.13 Ground surface data table of the synthetic dataset 3

x	y	z	priority_num	type	polarity	angle
80	80	10		Topography		
175	175	-4		Topography	Normal	
215	215	0	7	Topography	Normal	
219	219	0	8	Topography	Normal	
235	235	0	8	Topography	Normal	
239	239	0	7	Topography	Normal	
245	245	-4		Topography		

## C. Listings of Chapter 2

```

1. Structure Determination
2. priority: 3
3. priority_type conformity
4. #####
5. zone box:
6. | X | Y | Z |
7. |-----+-----|
8. | 90 | 90 | -10 |
9. | 130 | 130 | -10 |
10. | 130 | 130 | -55 |
11. | 90 | 90 | -55 |
12. ### RELATED FAULTS ###
13. fault priority: 6
14. fault points:
15.
16. | BOREHOLE_ID | [X,Y,Z] | POINT_ID |
17. |-----+-----|
18. | 112 | [90.0, 90.0, -47.0] | 39 |
19. | 113 | [110.0, 110.0, -20.0] | 41 |
20. ### ZONE INFORMATION ###
21.
22. | POINT_TYPE | BOREHOLEID|BOREHOLE_INDE| [X,Y,Z] | POLARITY | POINT SIT. | POINT_ID
23. |-----+-----+-----+-----+-----+-----+-----|
24. | Borehole_points | 111 | 2 | [70.0, 70.0, -50.0] | Normal | Not Connected | 3
25. | Borehole_points | 112 | 3 | [90.0, 90.0, -10.0] | Normal | Not Connected | 5
26. | Borehole_points | 112 | 3 | [90.0, 90.0, -25.0] | Reverse | Not Connected | 8
27. | Borehole_points | 112 | 3 | [90.0, 90.0, -55.0] | Normal | Not Connected | 9
28. | Borehole_points | 113 | 4 | [110.0, 110.0, -10.0] | Normal | Not Connected | 11
29. | Borehole_points | 113 | 4 | [110.0, 110.0, -35.0] | Reverse | Not Connected | 14
30. | Borehole_points | 113 | 4 | [110.0, 110.0, -55.0] | Normal | Not Connected | 15
31. | Borehole_points | 114 | 5 | [130.0, 130.0, -13.0] | Normal | Not Connected | 17
32. ##### Stage 2: Semi-Automatic #####
33. please introduce consecutive POINT_IDs that you wish to connect in form of a list.
34. In case the critical zone consists of more than one part, write 'SEPARATE' between POINT_IDs.
35. Note that the introduced points have to follow the borehole arrangement
36. In case this stage preferred to be done manually, Type:'jumptomannual'
37. EXAMPLE: [25,26,37,38,'SEPARATE',45,46,'SEPARATE',65,68] or 'jumptomannual'
38.
39. Enter the POINT_ID list:
40. [3,'SEPARATE',9,15,14,'SEPARATE',8,5,11,'SEPARATE',17]

```

Listing C.1 GEOPROPY semi-automatic (guided) stage of synthetic dataset 3

```

1. Structure Determination
2. priority: 4
3. priority_type intrusion
4. #####
5. zone box:
6. | X | Y | Z |
7. |-----|
8. | 180 | 180 | -38 |
9. | 230 | 230 | -38 |
10. | 230 | 230 | -63 |
11. | 180 | 180 | -63 |
12. ### RELATED FAULTS ###
13. No related fault detected
14. ### ZONE INFORMATION ###
15. | POINT_TYPE | BOREHOLEID|BOREHOLE_INDE| [X,Y,Z] | POLARITY | POINT SIT. | POINT_ID |
16. |-----|-----|-----|-----|-----|-----|-----|
17. | Borehole_points | 117 | 8 | [180.0, 180.0, -55.0] | Normal | Not Connected | 29 |
18. | Borehole_points | 117 | 8 | [180.0, 180.0, -63.0] | Normal | Not Connected | 30 |
19. | Borehole_points | 118 | 9 | [230.0, 230.0, -38.0] | Normal | Not Connected | 33 |
20. | Borehole_points | 118 | 9 | [230.0, 230.0, -63.0] | Normal | Not Connected | 34 |
21.
22. ##### Stage 2: Semi-Automatic #####
23. please introduce consecutive POINT_IDs that you wish to connect in form of a list.
24. In case the critical zone consists of more than one part, write 'SEPARATE' between POINT_IDs.
25. Note that the introduced points have to follow the borehole arrangement
26.
27. In case this stage preferred to be done manually, Type:'jumptomannual'
28.
29. EXAMPLE: [25,26,37,38,'SEPARATE',45,46,'SEPARATE',65,68] or 'jumptomannual'
30.
31. Enter the POINT_ID list:
32. 'jumptomannual'
33.
34. ##### Stage 3: Manual #####
35. There are points which are not connected from any side to any borehole. please specify how they have to
    connect. Use the instruction below:
36.
37. priority: 4
38. priority_type: intrusion
39.
40. ### ZONE INFORMATION ###
41.
42. | POINT_TYPE | BOREHOLEID|BOREHOLE_INDE| [X,Y,Z] | POLARITY | POINT SIT. | POINT_ID |
43. |-----|-----|-----|-----|-----|-----|-----|
44. | Borehole_points | 117 | 8 | [180.0, 180.0, -55.0] | Normal | Not Connected | 29 |
45. | Borehole_points | 117 | 8 | [180.0, 180.0, -63.0] | Normal | Not Connected | 30 |
46. | Borehole_points | 118 | 9 | [230.0, 230.0, -38.0] | Normal | Not Connected | 33 |
47. | Borehole_points | 118 | 9 | [230.0, 230.0, -63.0] | Normal | Not Connected | 34 |
48.
49. ##### 3rd STAGE GUIDE #####
50.
51. point IDs have to introduce in pairs, accompanied by a keyword that identify their situation:
52.
53. If both points are in 2 CONSECUTIVE borehole or surface points: [point_id 1, point_id 2, 'normal_connection']
    (point_id 1 is the sample point with smaller BOREHOLE_INDEX)
54.
55. If both points are in the same borehole, based on the side that you want them to connect: [point_id 1,
    point_id 2, 'same_bh_connect_left'] or [point_id 1, point_id 2, 'same_bh_connect_right']
56.
57. If it is preferred that the program take care of a point: [point_id 1, '', 'user_skipped'] (NOT RECOMMENDED)
58.
59. If it is preferred to introduce new coordinations and point ids, every group of lines (group of lines that are
    connected together) have to be introduced in different lists. the coordination of the new point and the
    related borehole id (if new introduced point is between two boreholes, use left side borehole id ) have to be
    introduced.
60. structure: [ [Coord_X,Coord_Y,Coord_Z,'BOREHOLE_ID' (string)], point_id1, point_id2,
    [Coord_X,Coord_Y,Coord_Z,'BOREHOLE_ID' (string)],... ] , '' , 'new_points'. Maintain sequence of the points.
    Note that in this part, it is possible to use points with ids that shown before.
61.
62. #####Disconnected sample points#####
63.
64. | POINT_TYPE | BOREHOLEID|BOREHOLE_INDE| [X,Y,Z] | POLARITY | POINT SIT. | POINT_ID |
65. |-----|-----|-----|-----|-----|-----|-----|
66. | Borehole_points | 117 | 8 | [180.0, 180.0, -55.0] | Normal | Not Connected | 29 |
67. | Borehole_points | 117 | 8 | [180.0, 180.0, -63.0] | Normal | Not Connected | 30 |
68. | Borehole_points | 118 | 9 | [230.0, 230.0, -38.0] | Normal | Not Connected | 33 |
69. | Borehole_points | 118 | 9 | [230.0, 230.0, -63.0] | Normal | Not Connected | 34 |
70. Enter POINT_ID list as mentioned:
71. [ [ [ 166,166, -71,'116'],29,33 ] ,'', 'new_points'], [ 33,34,'same_bh_connect_right'], [ [ 173,173,-75
    , '116'],30,34 ] ,'', 'new_points']]

```

Listing C.2 GEOPROPY third (manual) stage of synthetic dataset 3

#### **D. Urban cycle calculations – Chapter 4**

SWB calculations in urban area using the parameters defined in section 2.6 are as follows:

$$WSNLV = WSNC * WSNL$$

where WSNLV is Water Supply Network Loss Value.

$$WSNCV = WSNC - WSNLV$$

where WSNCV is Water Supply Network Consumption Value

$$WCNNLV = WCNN * WCNNL$$

where WCNNLV is Water Consumption NOT from network loss value.

$$WCNNV = WCNN - WCNNLV$$

where WCNNV is Water Consumption NOT from network value.

$$SNV1 = (WSNCV + WCNNV) * (1 - IUE)$$

where SNV1 is the first stage of sewage network value which is equal to the total network water consumption, not considering the indirect evaporation.

$$RtSV = (P + I) * (1 - DUE) * RtS$$

where RtSV is the runoff to sewage value

$$SNV2 = SNV1 + RtSV + WOS$$

where SNV2 is the second stage of sewage network value which adds the runoff to sewage value and the water from other sources that is directed to the sewage system.

$$SNLV2 = SNV2 * SNL$$

where SNLV2 is the second stage of sewage network loss value

$$SNV3 = SNV2 - SNLV2$$

where SNV3 is the third and final stage of sewage network value

$$DIV = (P + I) * DI$$

where DIV is the direct infiltration value.

$$\Delta WV = P + I + WSNC + WCNN + WOS - SNV3$$

where  $\Delta WV$  is equal to the difference in the value of all the water that enters the urban area minus the water that exists through the sewage network.

$$TIV = WSNLV + WCNNLV + SNLV2 + DIV$$

where TIV is the total infiltration value which is equal to the sum of water supply network loss value, Water Consumption NOT from network loss value, sewage network loss value and direct infiltration value.

$$TEPV = (WSNCV + WCNNV) * IUE + (P + I) * DUE$$

where TEPV is the total evaporation value which is the sum of direct and indirect evaporations.

$$TRV = \Delta WV - TETV - TIV$$

where TRV is the total runoff which is  $\Delta WV$  that is not infiltrated or evaporated in urban area.

## **E. Empirical methods equations and parameters – Chapter 4**

The Rao equations <sup>279</sup> in mm:

$$R = 0.3 * (P - 500)$$

The Chaturvedi <sup>276</sup> and modified Chaturvedi <sup>277</sup> equations respectively in inch:

$$R = 2 * (P - 15)^{0.4}$$

$$R = 1.35 * (P - 14)^{0.5}$$

The Maxey-Eakin <sup>278</sup> equation in mm:

$$R = A * P$$

The aforementioned relationships between P and R have to be adjusted based on the properties of the study area. In the example in section 6, the A in Maxey-Eakin has been considered equal to 0.1.



## F. Additional figures – Chapter 4

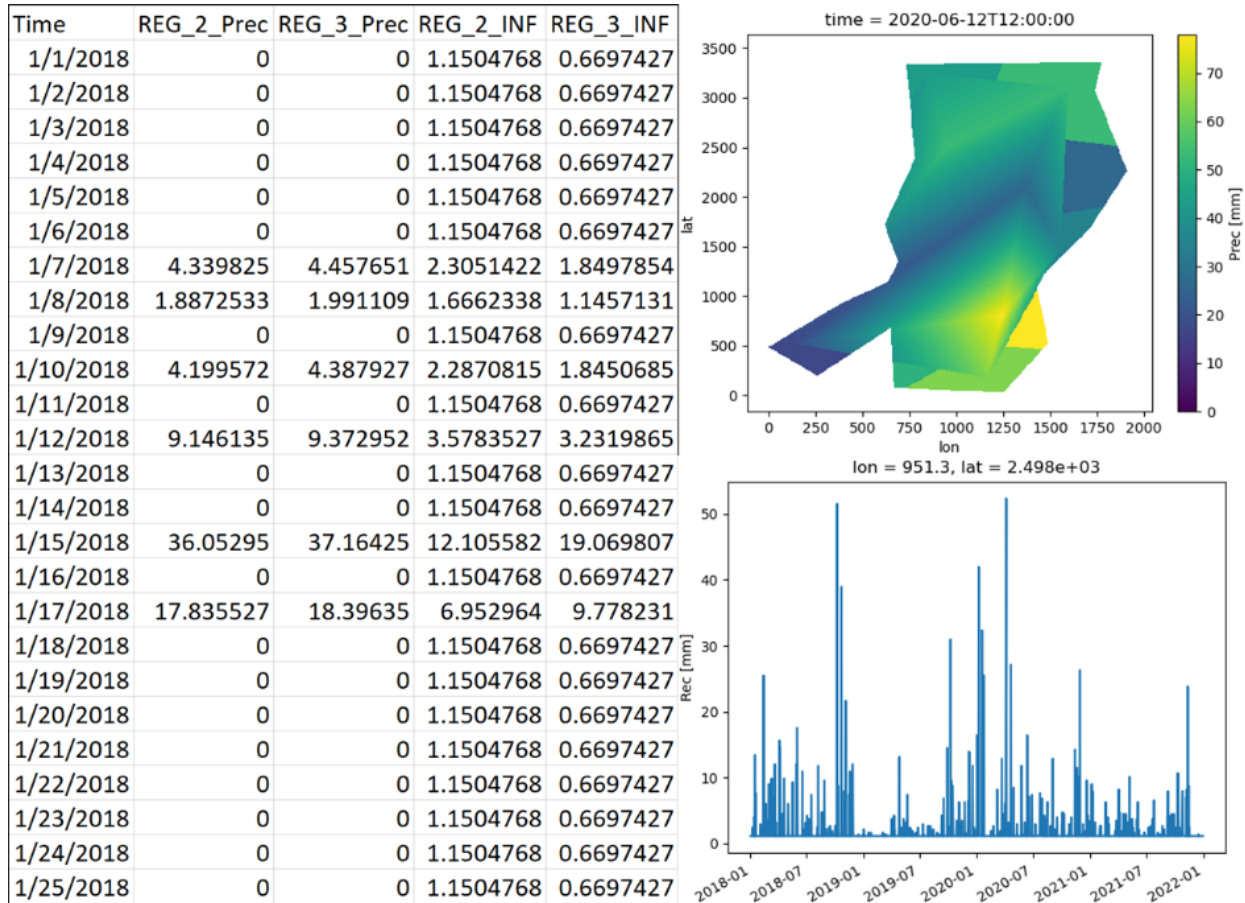


Figure F.1 The screenshots of 3 types of WateryBal output: Datasheets, map and time series chart.

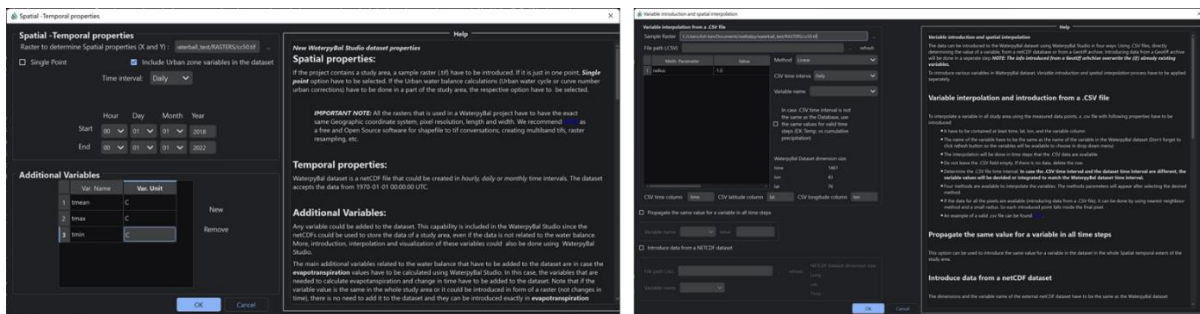


Figure F.2 The screenshots of WateryBal Studio. Left: The spatial-temporal properties window to define the properties of the dataset. Right: Variable introduction and interpolation window.

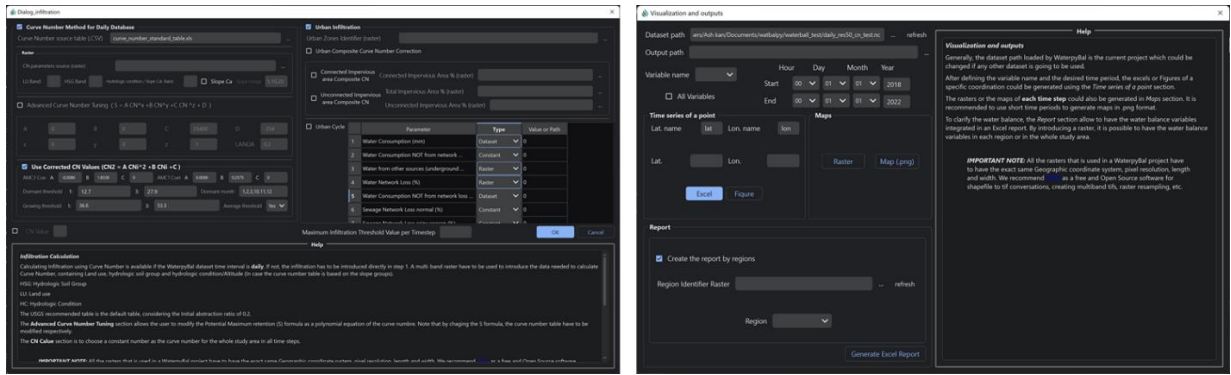


Figure F.3 The screenshots of WaterpyBal Studio. Left: Infiltration calculation window. Right: Visualization and output generation.

## G. List of Scientific and Technical Production

### Scientific articles

1. **Hassanzadeh, A.**, Vázquez-Suñé, E., Corbella, M. & Criollo, R. An automatic geological 3D cross-section generator: Geopropy, an open-source library. *Environmental Modelling & Software* 149, 105309 (2022). DOI: 10.1016/j.envsoft.2022.105309.
2. **Hassanzadeh, A.**, Valdivielso, S., Vázquez-Suñé, E., Criollo, R. & Corbella, M. An open source Python library for environmental isotopic modelling. *Scientific Reports* 2023 131 13, 1–19 (2023). DOI: 10.1038/s41598-023-29073-2.
3. **Hassanzadeh, A.**, Vázquez-Suñé, E., Valdivielso, S., Corbella, M. An open source Python library for water balance modelling. *Environmental Modelling & Software* (2023). (*submitted*)
4. Valdivielso, S., **Hassanzadeh, A.**, Vázquez-Suñé, E., Custodio, E. & Criollo, R. Spatial distribution of meteorological factors controlling stable isotopes in precipitation in Northern Chile. *Journal of Hydrology*. 605, 127380 (2022). DOI: 10.1016/j.jhydrol.2021.127380.
5. Valdivielso, S., Vázquez-Suñé, E., Lopez Moreno J.I., Custodio, E., Criollo, R., Pomeroy, J.W, **Hassanzadeh, A.**, Snowmelt contribution to groundwater recharge in the Salar de Atacama basin, Chile. *Journal of Hydrology* (2023). (*submitted*)

## Book chapters

E. Vázquez-Suñé, E. Queralt, M.J. Chesa, V. Solà, I. Bulboa, R. Criollo, J. Massana, F.J. Varela, S. Burdons, M. Enrich, L. Scheiber, **A. Hassanzadeh**, J. Botey, S. Valdivielso (2023). *HIDROGEOLOGIA DE L'ENTORN DE LA CIUTAT BARCELONA*. En: Vila, M. (coord.) 2022: *Geologia de la ciutat de Barcelona: coneixements, condicionants i aprofitaments d'una zona urbana complexa*, 388 pag. Col·lecció: Monografies tècniques 11, Institut Cartogràfic i Geològic de Catalunya, 2022. ISBN: 978-84-19326-88-1.

## Proceedings

1. Valdivielso, S., Vázquez-Suñé, E., **Hassanzadeh, A.** The Factors That Condition the Isotopic Composition of Rain in Northern Chile. in Goldschmidt Congress. 18-23 August 2019. Barcelona (Spain).

2. Valdivielso, S., **Hassanzadeh, A.** Drivers of Rainfall Isotope Composition in The Northern Chilean. in 46th International Association of Hydrogeologists (IAH) Congress. 22-27 September 2019. Malaga (Spain).

**3. Hassanzadeh, A.,** Vázquez-Suñé, E., Corbella, M. & Criollo, R. Geopropy: An open source tool to generate 3D geological cross sections. in EGU General Assembly. 23–27 May 2022. Vienna (Austria).  
doi:10.5194/egusphere-egu22-11486

4. **Hassanzadeh, A.,** Vázquez-Suñé, E., Valdivielso, S., Criollo, R. A water isotopic modelling library for environmental studies. in American Geophysical Union (AGU). 12 - 16 December 2022. Chicago (US).  
doi:10.22541/essoar.167590831.19426646/v1

5. Vázquez-Suñé, E., Valdivielso, S., **Hassanzadeh, A.**, Custodio, E. & Criollo, R. Factors that control the isotopic composition of precipitation in northern Chile. in EGU General Assembly. 23–27 May 2022. Vienna (Austria). doi:10.5194/egusphere-egu22-9346

6. **Hassanzadeh, A.**, Vázquez-Suñé, E., Una herramienta de código abierto para generar cortes geológicos en 3D: Geopropy. in International Association of Hydrogeologists (IAH) Congress. 23-25 November 2022. Albacete (Spain).

7. **Hassanzadeh, A.**, Valdivielso, S., Vázquez-Suñé, E., Criollo, R. & Corbella, M. An open source library for environmental isotopic modelling using machine learning techniques. in EGU General Assembly. 23–28 April 2023. Vienna (Austria). doi:10.5194/egusphere-egu23-1672

8. Valdivielso, S., Vázquez-Suñé, E., Lopez Moreno J.I., Custodio, E., Criollo, R., Pomeroy, J.W, **Hassanzadeh, A.** The importance of snowmelt in the water balance of the Toconao sub-basin, Salar de Atacama. in EGU General Assembly. 23–28 April 2023. Vienna (Austria). doi:10.5194/egusphere-egu23-315

### **Software Registration - Patents**

1. Geopropy Python Library. **Hassanzadeh, A.**, Vázquez-Suñé, E. Patent pending by CSIC

2. Isocompy Python Library. **Hassanzadeh, A.**, Vázquez-Suñé, E. Patent pending by CSIC

3. WaterpyBal Python Library. **Hassanzadeh, A.**, Vázquez-Suñé, E. Patent pending by CSIC

4. WaterpyBal Studio Computer Program **Hassanzadeh, A.**, Vázquez-Suñé, E. Patent pending by CSIC

## **Scientific and technology transfer projects**

1. **TOOLS AND CRITERIA FOR URBAN GROUNDWATER MANAGEMENT (URBANWAT)**, 2019 - 2022. ERA-NET Cofund WaterWorks (JPI Water), Grant agreement reference: PCI2019-103616.
2. **URBAN BIO-GEOCHEMISTRY: INTEGRATING THE AIR, WATER, SOIL AND MICROBIOLOGICAL SCIENCE NEEDED TO UNDERPIN POLLUTION MANAGEMENT. (UNBIASED)**. 2019 - 2021. Proyectos de I+D+i RETOS, AGENCIA ESTATAL DE INVESTIGACIÓN. Grant agreement reference: RTI2018-097346-B-I00
3. **AYUDAS EXTRAORDINARIAS PARA LA PREPARACIÓN DE PROYECTOS A REALIZAR EN EL MARCO DEL PLAN ESTATAL DE I+D+I (CSIC)**, January 2022 to September 2022. CONSEJO SUPERIOR DE INVESTIGACIONES CIENTIFICAS, Grant agreement reference:2021AEP052.
4. **UNDERSTANDING GROUNDWATER POLLUTION TO PROTECT AND ENHANCE WATER QUALITY (UPWATER)**, 2022 – 2026. UE - HORIZON-CL6-2022-ZEROPOLLUTION-01-01. Grant agreement reference: 101081807
5. **BUILDING URBAN WATER RESILIENCE: SUSTAINABLE URBAN DRAINAGE SYSTEMS TO TACKLE STORMWATER RUNOFF POLLUTION WHILE AUGMENTING AND RESTORING LOCAL WATER RESOURCES. (URBPOL)**, 2022 - 2024. Proyectos de Transición Ecológica y Transición Digital, AGENCIA ESTATAL DE INVESTIGACIÓN. Grant agreement reference: TED2021-132894B-I00.
6. **TRABAJOS DE MODELACIÓN HIDROGEOLÓGICA SQM-SALAR**. 2018 – 2021. SQM, Chile.
7. **ESTUDI I SEGUIMENT DE L'EVOLUCIÓ DEL DRENATGE I DELS NIVELLS FREÀTICS A L'ENTORN DE LA PLAÇA DE LA VILA DE SANT ADRIÀ DE BESÒS**. 2018 – 2022. Ajuntament de Sant Adrià del Besòs
8. **REVISIÓN CÁLCULOS HIDROGEOLÓGICOS Y SEGUIMIENTO DE LOS EFECTOS PRODUCIDOS POR EL AGOTAMIENTO DEL FREÁTICO EN EL ÁMBITO DE LA**

**EJECUCIÓN DEL TÚNEL DE LA PLAZA DE LAS GLORIAS.** 2018 – 2021. Ajuntament de Barcelona / BIMSA

9. **APOYO TECNOLÓGICO PARA LA VALIDACIÓN, SISTEMATIZACIÓN Y APOYO A LA IMPLEMENTACIÓN DE UN MODELO HIDROLÓGICO E HIDROGEOLÓGICO DE LA CUENCA KATARI Y LAGO MENOR DEL TITICACA (BOLIVIA).** September 2019 – November 2019. Banco Interamericano de Desarrollo (BID)
10. **INTEGRACIÓ DE DADES GEOLÒGIQUES I HIDROLÒGIQUES, I VALIDACIÓ I CONSOLIDACIÓ DE LES DADES PRÈVIES DE LA NAU 8 DEL CONJUNT INDRUSTRIAL DE CAN BATLLÓ, AL DISTRICTE DE SANTS-MONTJUÏC, A BARCELONA.** October 2019 – December 2019. Ajuntament de
11. **MODELADO DEL FLUJO SUBTERRÁNEO EN EL ÁMBITO DEL TRAZADO DE LA LÍNEA 8 Y EVALUACIÓN DE EFECTOS SOBRE LOS ACUÍFEROS DE LA CIUDAD DE BARCELONA.** 2020 – 2021. UTE VALLOAN
12. **TRABAJOS DE CARACTERIZACIÓN HIDROGEOLÓGICA E HIDROGEOQUÍMICA EN RELACIÓN AL AGOTACIÓN DEL FREÁTICO EN AVDA. EDUARD MARISTANY, PARCELA 12, BADALONA.** June 2021 – August 2021. PREMIER PROYECTOS Y PORMOCION DE VIVIENDAS S.L.
13. **HYDROGEOLOGICAL INVESTIGATIONS FOR THE DESIGN OF SOUTH OF WAKRAH STORMWATER TUNNELS, PUMPING STATION AND TUNNEL OUTFALL, QATAR.** November 2021 – May 2022. Stantec UK Ltd - Qatar Branch
14. **REVISIÓN DE ESTUDIOS HIDROGEOLÓGICOS EN DIFERENTES ÁREAS DE INTERÉS PARA CODELCO (CHILE).** 2022 – 2024. CORPORACION NACIONAL DEL COBRE DE CHILE (CODELCO)
15. **ACTUALITZACIÓ GEOLÒGICA E HIDROGEOLÒGICA DEL MARGEN ESTE DEL DELTA DEL LLOBREGAT (SECTOR PUERTO DE BARCELONA).** 2022 – 2023. Comunitat d'Usuaris del Delta del Llobregat (CUADLL)

## H. Cover of the scientific articles

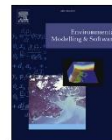
Environmental Modelling and Software 149 (2022) 105309



Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: [www.elsevier.com/locate/envsoft](http://www.elsevier.com/locate/envsoft)



### An automatic geological 3D cross-section generator: Geopropy, an open-source library

Ashkan Hassanzadeh<sup>a,b,\*</sup>, Enric Vázquez-Suñé<sup>a</sup>, Mercè Corbella<sup>b</sup>, Rotman Criollo<sup>c</sup>

<sup>a</sup> Institute of Environmental Assessment and Water Research (IDAEA-CSIC), C/ Jordi Girona 18-26, 08034, Barcelona, Spain

<sup>b</sup> Universitat Autònoma de Barcelona (UAB), Departament de Geologia, Edifici C, 08193, Bellaterra, Barcelona, Spain

<sup>c</sup> Mediterranean Institute for Advanced Studies (IMEDEA, CSIC-UIB), C/ Miquel Marqués, 21, 07190, Esporles, Illes Balears, Spain

#### ARTICLE INFO

##### Keywords:

Decision making algorithm  
3D geological modelling  
Cross-section  
Open source  
Python

#### ABSTRACT

Geological modelling is an essential aspect of underground investigations, with cross-sections being one of the key aspects. This modelling can be done by experienced geologists or using mathematical methods. We present Geopropy, an open-source decision-making algorithm implemented in Python, that generates 3D cross-sections (the boreholes do not have to be aligned). It performs as an intelligent agent that simulates the steps taken by the geologist in the process of creating the cross-section, coupled with data-driven decisions. The algorithm detects zones with more than one possible outcome and, based on the level of complexity (or user preference), proceeds to automatic, semiautomatic or manual stages. Geopropy could be the basis of a new, simpler, more comprehensible way of looking at geological models in industry and academia while at the same time creating the potential for using novel machine learning algorithms in geological modelling.

#### 1. Introduction

Obtaining 3D subsurface geological models is of great importance in a wide variety of geoscience studies. These models represent superficial and underground geological structures and the distribution of geological units, and their purpose is to illustrate the existing underground conditions. Subsurface geological understanding and models are essential aspects of earth-related industrial projects and academic investigations, from mining and petroleum to hydrogeology and environmental studies (Hawie et al., 2021; Kerrou et al., 2017; Pasculli et al., 2014; Yang et al., 2021).

In general, geological models can be described from implicit or explicit points of view. Implicit modelling includes data-driven methods that use datasets derived from measured features and algorithms. Explicit modelling mostly relies on expert opinion, experience and interpretation, where the expert usually builds cross-sections, surfaces or volumes by interpolating the accessible data (Randle et al., 2018).

To date, various works have presented and discussed the use of implicit methods, including cokriging-based modelling (Calcagno et al., 2008; Gonçalves et al., 2017; Hillier et al., 2014; Lajaunie et al., 1997), volume-based modelling (Iskenova et al., 2016; Souche et al., 2013), kinematic modelling (Brandes and Tanner, 2014; Caumon, 2010) and

others (Lemon and Jones, 2003; Muzik et al., 2015). In addition, open-source software such as Noddy/pynoddy (Florian Wellmann et al., 2016; Jessell and Valenta, 1996) and GemPy (De La Varga et al., 2019; Schaaf et al., 2020) and commercial software such as GDM Suite (BRGM, 2020), MOVE (Petroleum Experts, 2019), Vulcan 3D/EUREKA (Maptek, 2021), Leapfrog (Seequent, 2021a) and Oasis Montaj (Seequent, 2021b) are powered by these methods. Most of these computer programs contain more than one method, and although they are known as implicit modelers, they benefit from explicit modelling in some stages of geological model evaluation.

For explicit modelling, there are computer programs that support experts in generating cross-sections. Some examples of open-source software are HEROS (Velasco et al., 2013) and HEROS 3D (Alcaraz et al., 2016b), the Midvatten QGIS plugin (Källgård and Spångmyr, 2020), Grass GIS (GRASS Development Team, 2020), and Geomodelr (Serrano, 2019). There are also commercial software programs that support explicit modelling, such as RockWare GIS (RockWare, 2020), Surpac (Dassault Systèmes, 2021), Datamine (2021), GeoScene3D (I-GIS, 2020), BGS Groundhog Desktop (British Geological Survey, 2020), and GSI3D (Cullen et al., 2010). One can also use other characteristics to categorize these programs: the amount or type of information used to build a model, the proportion of knowledge or data that drives

\* Corresponding author. Institute of Environmental Assessment and Water Research (IDAEA-CSIC), C/ Jordi Girona 18-26, 08034, Barcelona, Spain.  
E-mail address: [ashkan.hassanzadeh@csic.es](mailto:ashkan.hassanzadeh@csic.es) (A. Hassanzadeh).

<https://doi.org/10.1016/j.envsoft.2022.105309>

Received 26 April 2021; Received in revised form 20 December 2021; Accepted 6 January 2022

Available online 13 January 2022


1364-8152/© 2022 The Authors. 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/>).





# OPEN An open source Python library for environmental isotopic modelling

Ashkan Hassanzadeh<sup>1,2</sup>, Sonia Valdivielso<sup>1</sup>, Enric Vázquez-Suñé<sup>1</sup>, Rotman Criollo<sup>3</sup> & Mercè Corbella<sup>2</sup>

Isotopic composition modelling is a key aspect in many environmental studies. This work presents *Isocompy*, an open source Python library that estimates isotopic compositions through machine learning algorithms with user-defined variables. *Isocompy* includes dataset preprocessing, outlier detection, statistical analysis, feature selection, model validation and calibration and postprocessing. This tool has the flexibility to operate with discontinuous inputs in time and space. The automatic decision-making procedures are knitted in different stages of the algorithm, although it is possible to manually complete each step. The extensive output reports, figures and maps generated by *Isocompy* facilitate the comprehension of stable water isotope studies. The functionality of *Isocompy* is demonstrated with an application example involving the meteorological features and isotopic composition of precipitation in N Chile, which are compared with the results produced in previous studies. In essence, *Isocompy* offers an open source foundation for isotopic studies that ensures reproducible research in environmental fields.

Water isotopic composition is of paramount importance for decision making in many fields of study, including environmental resource management<sup>1</sup>. The stable water isotopes <sup>18</sup>O and <sup>2</sup>H are indicators of diverse aspects of the hydrological cycle.  $\delta^{18}\text{O}$  and  $\delta^2\text{H}$  measurements in precipitation are utilized in different meteorological and hydrological studies to identify the origin of precipitation, recognize local effects in water cycle studies, define the relative shares of water with different origins in a water body, describe aquifer recharging and characterization process and investigate various aspects of runoff and stream flow generation. All these features are essential for the optimal and sustainable management of water resources<sup>2,3</sup>.

The isotopic composition of rainwater is influenced by different physical variables and processes: temperature; pressure; humidity during condensation (to generate precipitation)<sup>4,5</sup>; mixtures of air masses with distinct origins<sup>6</sup>; the isotopic composition of the seawater from which air moisture condenses<sup>7</sup>; in-cloud microphysical processes<sup>8–12</sup>; the moisture conditions below clouds and the partial evaporation of precipitation along the path between clouds and the ground<sup>13–15</sup>; and the mixture of recycled precipitation from evapotranspiration over continents<sup>16–18</sup>. Therefore, detailed isotopic signature studies are used to discern these effects in any study area.

A linear relationship called the global meteoric water line (GMWL) is present between the  $\delta^{18}\text{O}$  and  $\delta^2\text{H}$  of meteoric water at the global scale, and this relationship is defined as  $\delta^{18}\text{O} = 8 \times \delta^2\text{H} + 10$ <sup>14</sup>. The characteristic isotopic signature of meteoric water in a particular region is caused by the various temperatures, relative humidity values, amounts of precipitation, latitudes and landmass proximities. The water molecules components (O, H) undergo isotope fractionation during phase transitions and the ratios of heavy versus light isotopes acts as a traceable feature of the physical processes<sup>19–23</sup>.

Two common approaches are available for studying the global distribution of the isotopic composition of precipitation: isotope-enabled atmospheric general circulation models (IGCMs) and regression statistics-based approaches<sup>24</sup>. IGCMs are numerical models that improve our understanding and reveal valuable information of the atmosphere by considering different physical processes (diffusion, advection, convection, etc.), including the physics of water isotopes (e.g., isotope fractionation, evaporation, condensation, among others)<sup>25</sup>. Computational power and numerical modelling advancements in recent decades have played an important role in the development of IGCMs, as they have resulted in a variety of models at different regional scales with diverse levels of complexity, such as CAM5<sup>26–28</sup>, ECHAM5<sup>29,30</sup>, MIROC<sup>31</sup> and LMDZ4<sup>32</sup>. IGCMs are usually complex,

<sup>1</sup>Institute of Environmental Assessment and Water Research (IDAEA/CSIC), C/ Jordi Girona 18-26, 08034 Barcelona, Spain. <sup>2</sup>Departament de Geologia, Universitat Autònoma de Barcelona (UAB), Edificis C, Bellaterra, 08193 Barcelona, Spain. <sup>3</sup>Mediterranean Institute for Advanced Studies (IMEDEA, UIB-CSIC), 07190 Esporles, Spain. ✉email: ashkan.hassanzadeh@csic.es

# An open-source Python library for water balance modeling

Ashkan Hassanzadeh<sup>1,2\*</sup>, Eric Vázquez-Suñé<sup>1</sup>, Sonia Valdivielso<sup>1</sup>, Mercè Corbella<sup>2</sup>

<sup>1</sup> Institute of Environmental Assessment and Water Research (IDAEA/CSIC), C/ Jordi Girona 18-26, 08034, Barcelona, Spain

<sup>2</sup> Universitat Autònoma de Barcelona (UAB), Departament de Geologia, Edificis C, 08193, Bellaterra, Barcelona, Spain

## Abstract

Water balance modeling is an essential aspect of water management projects. This study presents WaterpyBal, a tool that generates spatial–temporal water balance models, focused on diffused precipitation and recharge modeling. WaterpyBal has flexible input data and modeling time intervals and integrates different stages of the water balance assessment, such as data interpolation and evapotranspiration and infiltration calculations, while taking into account soil characteristics and urban water cycle. WaterpyBal calculates water budget parameters such as recharge, deficit, and runoff and can be used to create maps, dashboards, and raster archives. WaterpyBal has a modular design that ensures its further development. This Python library is accompanied by the WaterpyBal Studio, a graphic user interface that facilitates the usage of WaterpyBal in water management projects. The functionality of WaterpyBal is demonstrated using a synthetic example. In essence, WaterpyBal supports sustainable groundwater management projects and ensures reproducible research results in these environmental fields.

**Keywords:** water balance modelling, soil water balance, python, urban water cycle

## 1 Introduction

As the global population continues to grow and climate change intensifies, sustainable groundwater management becomes increasingly crucial for ensuring long-term access to safe and reliable water resources (Verma et al., 2021). One of the key aspects of groundwater management is groundwater recharge assessments, which quantify precipitation and other water resources that enter groundwater reservoirs (aquifers). However, many methods have been used to estimate recharge, choosing the appropriate method is of great importance (Weatherl et al., 2021). A common approach for recharge estimation is based on the source of information employed, such as surface water, unsaturated and saturated zone techniques, empirical formulas, or a mix of these methods (Scanlon et al., 2002).

Surface water–based approaches generally focus on the relationship between the aquifer and surface water dynamics, which is determined by soil characteristics (Sophocleous, 2002). Unsaturated zone–based techniques estimate groundwater recharge based on the drainage below the root zone (Custodio, 2019; Hauwert et al., 2014; Simmers et al., 2017; Vinet and Zhedanov, 2011). Recharge estimations using saturated zone–based techniques are generally derived from observed data in saturated zones such as the groundwater level measurements (Healy and Cook, 2002).

Advancements in computation power and ability have resulted in the design of many computer programs that can aid experts in recharge assessment. The programs for recharge assessment are based on different methods and useful in different manners, but these programs have some shortcomings that have not been fully addressed to date:

1. Homogeneity of spatial data. Depending on the objectives of the study and data availability, groundwater recharge assessments vary greatly in scale, from regional to local studies (Custodio, 2019). Some programs have limited spatial variability (Easy Bal) (Serrano-Juan et al., 2020).
2. The time window. Different time windows (hourly, daily, monthly, etc.) have to be used depending on the objectives of the study, data availability, and so forth. Computer programs bound by a specific time window can be potentially limited in this aspect (SWAT) (Arnold et al., 1998), (HYDROBAL) (Bellot and Chirino, 2013), (SWB) (U.S. Geological Survey, 2019), (HydroBudget) (Dubois et al., 2021).



Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: [www.elsevier.com/locate/jhydrol](http://www.elsevier.com/locate/jhydrol)

## Spatial distribution of meteorological factors controlling stable isotopes in precipitation in Northern Chile

Sonia Valdivielso<sup>a,b,\*</sup>, Ashkan Hassanzadeh<sup>a,c</sup>, Enric Vázquez-Suñé<sup>a</sup>, Emilio Custodio<sup>d,e,f</sup>, Rotman Criollo<sup>g</sup>

<sup>a</sup> Institute of Environmental Assessment and Water Research (IDAEA-CSIC), C/ Jordi Girona 18-26, 08034 Barcelona, Spain

<sup>b</sup> University of Barcelona (UB), C/ Martí i Franqués, 08028 Barcelona, Spain

<sup>c</sup> Autonomous University of Barcelona (UAB), Cívica Square, 08193 Bellaterra Barcelona, Spain

<sup>d</sup> Groundwater Hydrology Group, Department of Civil and Environmental Engineering, Technical University of Catalonia (UPC), C/ Jordi Girona 1-3, 08034 Barcelona, Spain

<sup>e</sup> Royal Academy of Mathematical, Physical and Natural Sciences (RAC) of Spain, Spain

<sup>f</sup> Instituto de Estudios Ambientales y Recursos Naturales (IUNAT), University of Las Palmas de Gran Canaria, Islas Canarias, Spain

<sup>g</sup> Mediterranean Institute for Advanced Studies (IMEDEA, CSIC-UIB). C/Miquel Marqués, 21, 07190 Esporles - Illes Balears, Spain

### ARTICLE INFO

This manuscript was handled by Corrado Corradini, Editor-in-Chief, with the assistance of Frédéric Juneau, Associate Editor

#### Keywords:

Central Andes  
Northern Chile  
Spatial distribution  
Climatology  
Isotopic composition  
Precipitation  
Water resources generation

### ABSTRACT

A knowledge of the evolution of isotopic composition of air masses humidity and precipitation in the Western Cordilleras of the Central Andes is still incomplete. This study contributes to a better understanding of the factors that control the  $\delta^{18}\text{O}$  and  $\delta^2\text{H}$  contents in precipitation water in the north of Chile, above 2000 m a.s.l. This paper deals with: (1) the relevant effects and processes that control the spatial (longitude, latitude and altitude) distribution of stable isotope contents of precipitation events in northern Chile, (2) the influence of local meteorological variables: temperature, precipitation and relative humidity on the  $\delta^{18}\text{O}$  and  $\delta^2\text{H}$  of precipitation, and (3) the estimation of these meteorological and isotopic variables at specific sites. To achieve these objectives, the relationships between geospatial and meteorological values are identified and analysed, followed by the estimation with empirical models. These estimation models (linear and non-linear) are obtained after examining, validating and calibrating techniques to find the best fit. This results in models for temperature, relative humidity and precipitation for each month of the year. In the same way, three isotopic models are derived from the spatial and meteorological variables (summer, winter and annual). Temperature has been shown to be controlled to a greater extent by altitude and latitude, while relative humidity is by latitude and precipitation in summer is by altitude and latitude. Monthly meteorological variables have been estimated throughout the study area. Precipitation  $\delta^{18}\text{O}$  and  $\delta^2\text{H}$  are controlled mainly by temperature and altitude and to a lesser extent by latitude, longitude and precipitation. In the same way, three isotopic models are derived from the spatial and meteorological variables: summer, winter and annual. This opens a new perspective of precipitation and its isotopic contents, but also allows the calculation of runoff and aquifer recharge and the path for linking future precipitation and aquifer recharge through their isotopic composition.

## 1. Introduction

### 1.1. General considerations

The water resources of a given area depend on precipitation, in addition to water coming from neighbouring territories. Part of precipitation produces aquifer recharge through diffuse processes, and infiltration of runoff and other surface water in parts of the territory. The

scarce water resources in arid areas generally have important environmental and ecological roles and support traditional local human populations, but they are also needed to supply other beneficial human activities, such as agriculture, raising animals, urbanization, mining, industry and tourism. These demands all compete among themselves for the available water resources. Consequently, management is needed, as well as setting the basis for sound water governance. Decision-making needs the support of different types of studies, as well as the

\* Corresponding author.

E-mail address: [sonia.valdivielso@idaea.csic.es](mailto:sonia.valdivielso@idaea.csic.es) (S. Valdivielso).

<https://doi.org/10.1016/j.jhydrol.2021.127380>

Received 25 August 2021; Received in revised form 18 November 2021; Accepted 17 December 2021

Available online 24 December 2021

0022-1694/© 2021 Elsevier B.V. All rights reserved.



This thesis is based on open-source tools to aid in the different stages of groundwater studies: geological 3D modeling (Geopropy), isotopic modeling (Isocompy), and soil water balance spatial-temporal modeling (WaterpyBal and WaterpyBal Studio). Geopropy is a decision-making algorithm implemented in Python that generates 3D cross-sections. It performs as an intelligent agent that simulates the steps taken by the geologist in the process of creating the cross-section, coupled with data-driven decisions. The algorithm detects zones with more than one possible outcome and, based on the level of complexity (or user preference), proceeds to automatic, semiautomatic or manual stages. Isocompy is Python library that estimates isotopic compositions through machine learning algorithms with user-defined variables. It includes dataset preprocessing, outlier detection, statistical analysis, feature selection, model validation and calibration and postprocessing. The automatic decision-making procedures are knitted in different stages of the algorithm, although it is possible to manually complete each step. WaterpyBal is another tool implemented in Python that generates spatial-temporal water balance models. WaterpyBal focuses on diffused precipitation and recharge modelling, considering the vertical water movement. This tool integrates different stages of the water balance assessment such as spatial data interpolation, evaporation, evapotranspiration and infiltration calculation, taking into account the soil characteristics and urban water cycle parameters. WaterpyBal calculates the water budget parameters such as recharge, deficit and runoff in defined spatial-temporal spectrum. WaterpyBal Studio is the graphic user interface of WaterpyBal that facilitates the usage of this tool in water management projects. In essence, these tools offer support for sustainable groundwater management projects that ensures reproducible research results in these environmental fields.

## Advisors

Dr. Enric Vázquez-Suñé

Dr. Mercè Corbella

