




OPEN

An open source Python library for environmental isotopic modelling

Ashkan Hassanzadeh^{1,2}, Sonia Valdivielso¹, Enric Vázquez-Suñé¹, Rotman Criollo³ & Mercè Corbella²

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 management¹. The stable water isotopes ¹⁸O and ²H 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^{2,3}.

The isotopic composition of rainwater is influenced by different physical variables and processes: temperature; pressure; humidity during condensation (to generate precipitation)^{4,5}; mixtures of air masses with distinct origins⁶; the isotopic composition of the seawater from which air moisture condenses⁷; in-cloud microphysical processes^{8–12}; the moisture conditions below clouds and the partial evaporation of precipitation along the path between clouds and the ground^{13–15}; and the mixture of recycled precipitation from evapotranspiration over continents^{16–18}. 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 $\delta^{18}\text{O} = 8 \times \delta^2\text{H} + 10$ ¹⁴. 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^{19–23}.

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 approaches²⁴. 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)²⁵. 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^{26–28}, ECHAM5^{29,30}, MIROC³¹ and LMDZ4³². IGCMs are usually complex,

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

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 results²⁵.

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^{24,33}. 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 aspects³⁴. 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 (“[Methods](#)”) and the different aspects of Isocompy (“[Under the hood of Isocompy](#)”), and we demonstrate its functionality by applying it to an example involving Salar de Atacama (Chile) (“[Application to the example of Salar de Atacama](#)”).

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^{35–41}. However, other information may be needed, such as sea surface temperatures, atmospheric pressures, outgoing longwave radiation (OLR) values^{10,42–45}, features derived from air mass trajectories^{39,46} or features resulting from reanalysis (such as wind components, dewpoint temperatures, and evaporation values)⁴⁷. 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. Figure 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. Figure 1(2,3) 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).

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 10- and 6-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

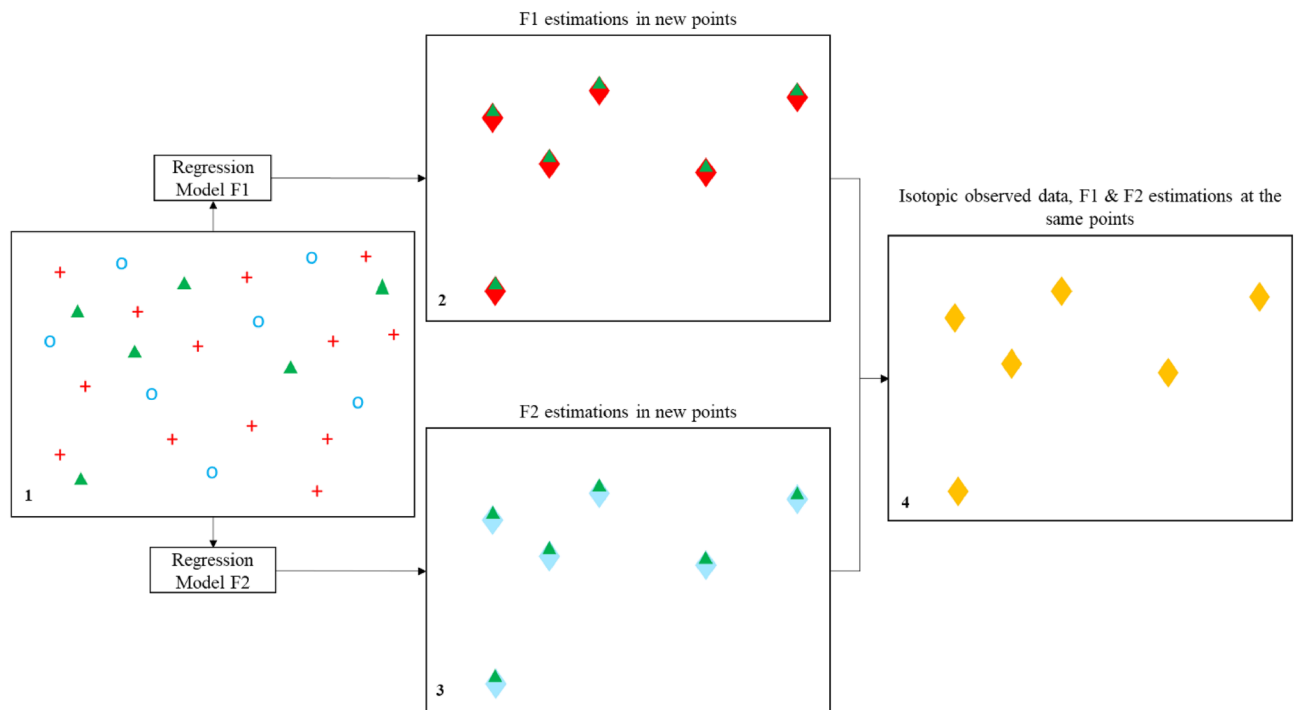


Figure 1. Workflow scheme for estimating isotopic values by using two independent parameters that are available in different locations than the isotopic measurements. Blue circles and red crosses represent two independent features that are potential candidates affecting the green triangles (isotopic 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, 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.

measured directly alongside the isotopic water composition or estimated via other methods, such as features derived from reanalysis^{48,49}. 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: Elasticnet⁵⁰, Bayesian ridge regression⁵¹, least-angle regression⁵², Bayesian automatic relevance determination (ARD)⁵³ and orthogonal matching pursuit⁵⁴, support vector regression⁵⁵, a random forest⁵⁶, and a multilayer perceptron⁵⁷. 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 process⁵⁸. The brute-force hyperparameter search algorithm is used to obtain a suitable set of hyperparameters⁵⁹; 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. Other data transformation techniques can be applied on the input data by the user.

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 in⁵⁰.

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 errors⁵⁴.

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 data⁵².

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 regression⁵¹.

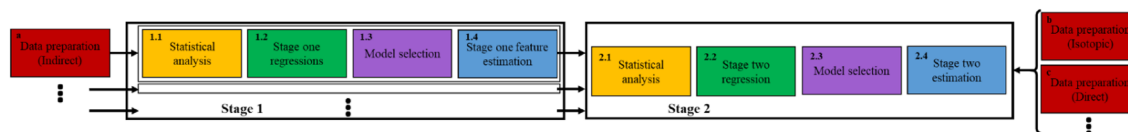


Figure 2. Scheme of the Isocompy workflow utilized to design the Isocompy architecture. It consists of data preparation (red boxes) and two main stages. Each stage includes statistical analysis (yellow boxes 1.1 and 2.1), regression (green boxes 1.2 and 2.2), model selection (violet boxes 1.3 and 2.3) and feature estimation steps (blue boxes 1.4 and 2.4).

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 estimates⁵⁶.

Support vector machines are versatile supervised learning methods that are used in various environmental science fields⁶⁰. 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 demanding⁶¹. Moreover, if the number of features is higher than the number of samples, the seed functions must be selected in a way that avoids overfitting⁶².

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 technique⁵⁷. However, an MLP is complex during the process of choosing the correct estimator hyperparameters.

Under the hood of Isocompy

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,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.

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.

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 model⁶³. 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 information⁶⁴, 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^{65,66}. 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

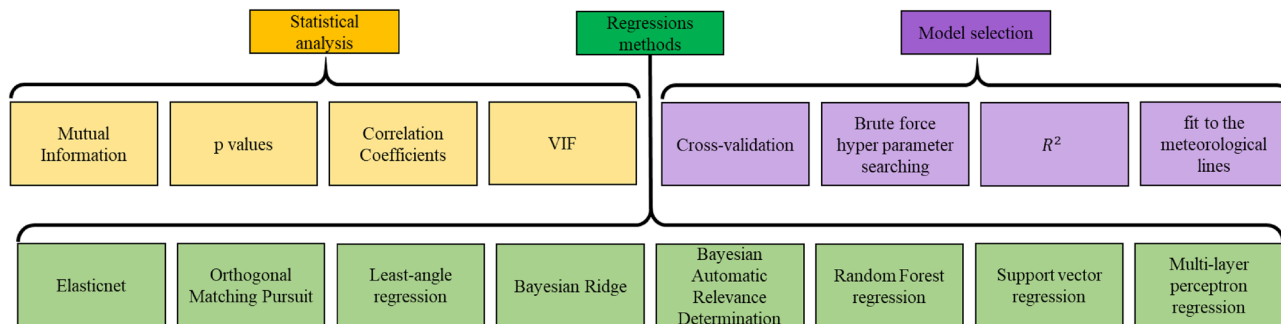


Figure 3. Yellow, green and violet boxes show the techniques used in the statistical analysis step, the regression methods available in Isocompy and the implemented techniques in the model selection steps, respectively.

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 “Methods”. 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.

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 pandas⁶⁷, geopandas⁶⁸, numpy⁶⁹, pylr²⁷⁰, statsmodels⁷¹ and scikit_learn⁷². 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 “[Isocompy library information](#)”.

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^{73,74}. 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

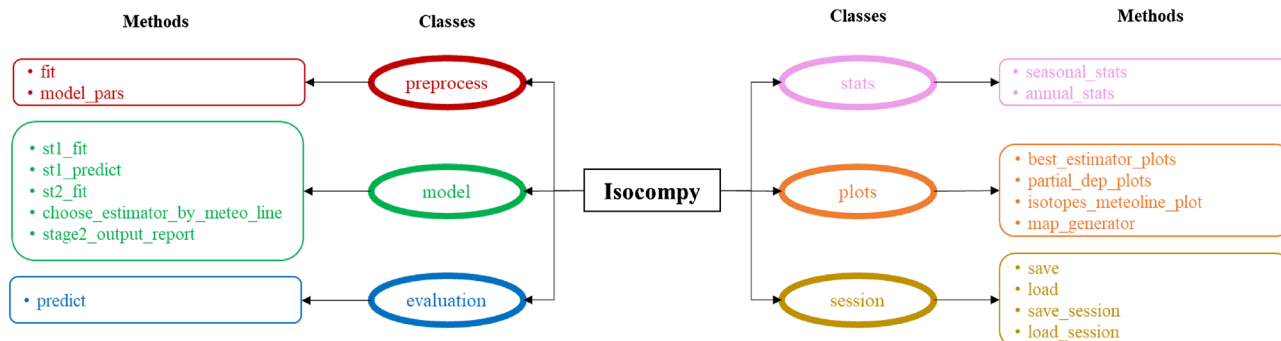


Figure 4. The Isocompy algorithm architecture. It contains 6 classes and 18 methods.

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 dataset⁷⁵. 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 “Isocompy library information”, 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 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.

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.

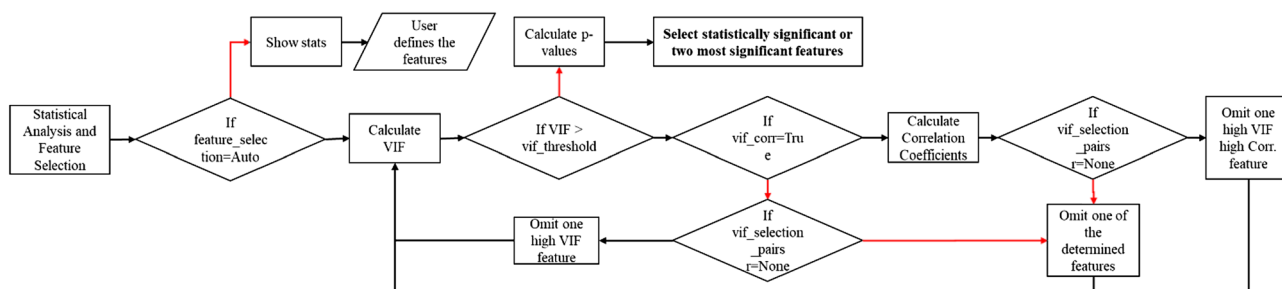


Figure 5. Feature selection flowchart of the model class. Red lines indicate false arguments.

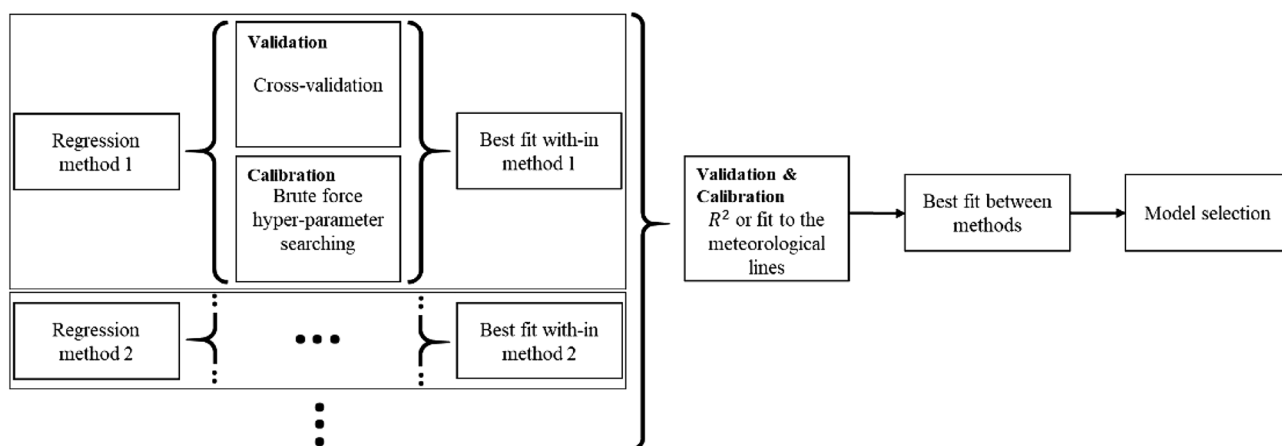


Figure 6. Workflow of the model regression, model validation, model calibration and best model selection processes. Black dots show that these processes are performed for each regression method selected.

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.

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*⁷⁶, *Bokeh*⁷⁷ and *matplotlib*⁷⁸ 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^{79–81}. 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 library⁸² 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.

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 “Postprocessing tools”. 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 “*Isocompy* architecture”. 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*.

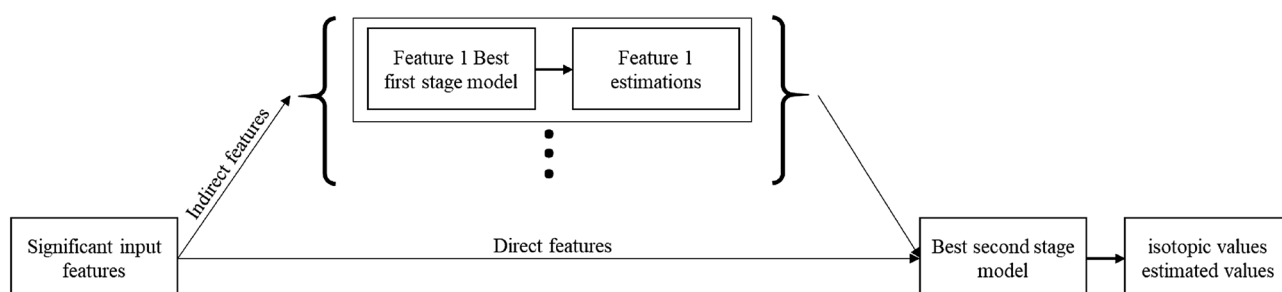


Figure 7. Workflow of the evaluation class for estimating the second-stage regressions.

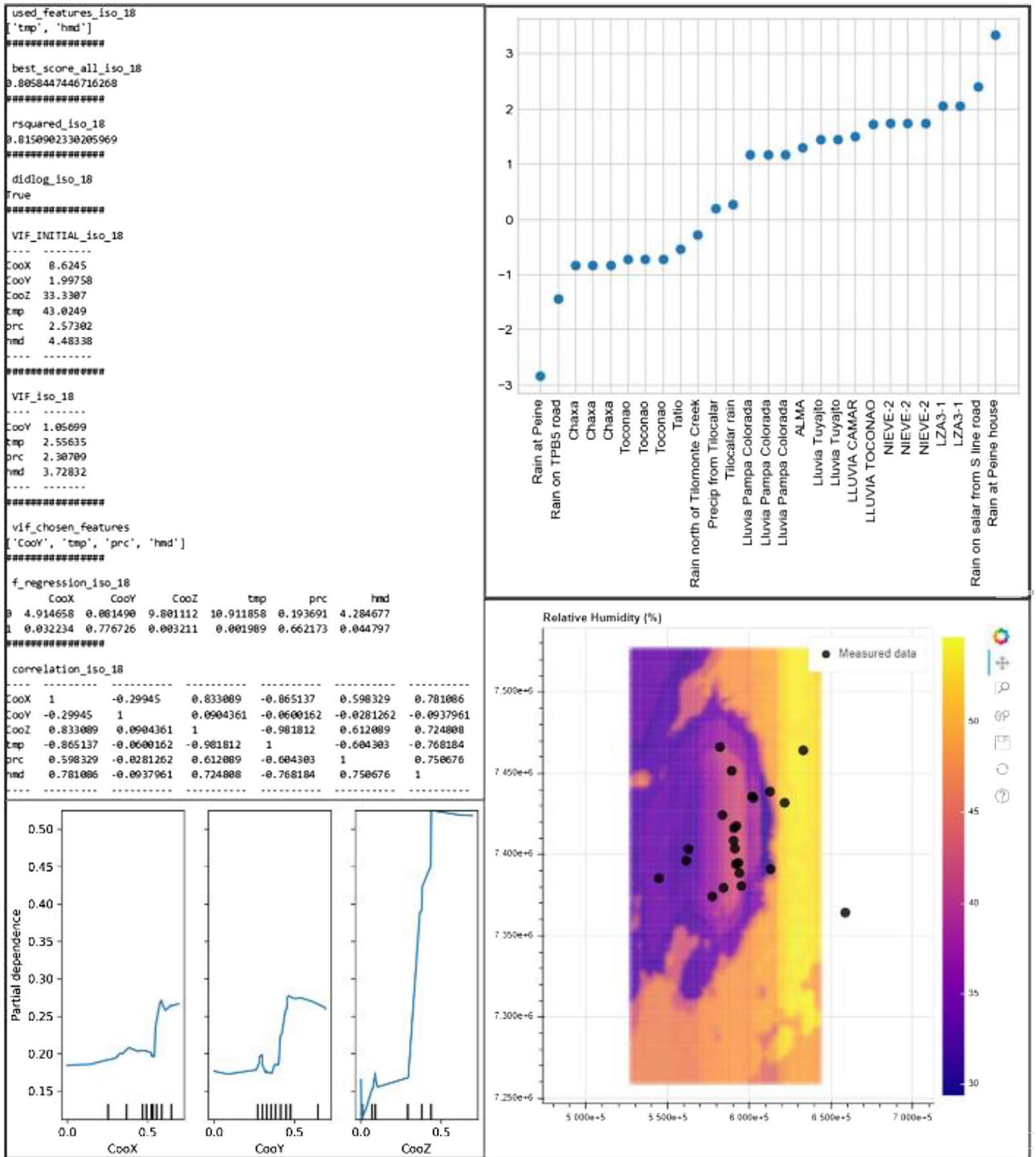


Figure 8. Screenshots of the outputs generated by Isocompy. Top left: an example report. Bottom left: partial dependency plots of the selected features. The values are standardized between zero and one. Vertical ticks on the x-axis illustrate the percentile of the data. Top right: a residual plot at 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.

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 “[Application on Salar de Atacama](#)” for an example datasheet.

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^{83–90}, 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^{83–90}. 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 stations⁹¹. 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.

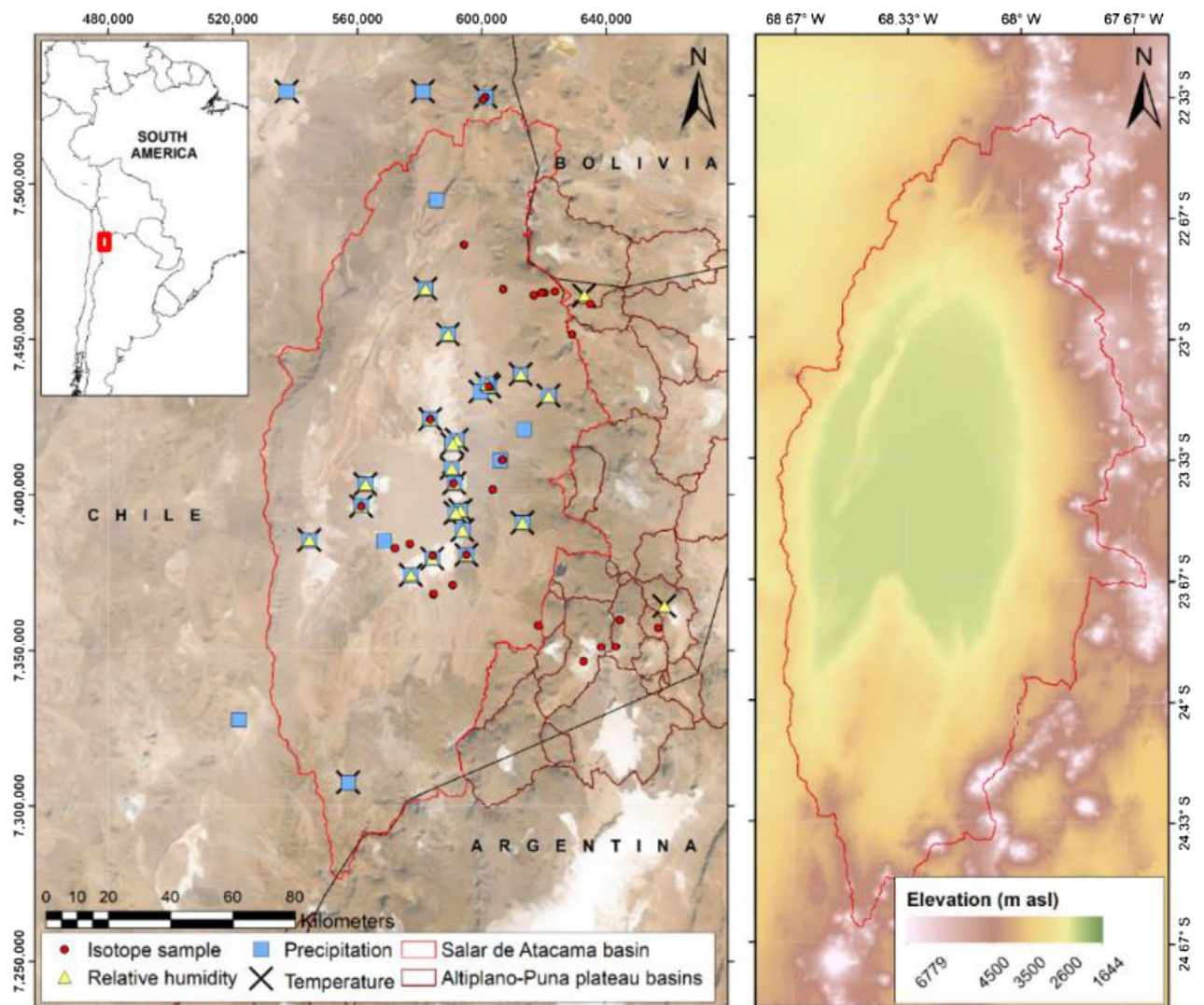


Figure 9. Left: location map of the study area in South America with published isotopic precipitation data (red circles) and automatic weather stations that monitor temperature (crosses), precipitation levels (blue squares) and relative humidity (yellow triangles). The solid red line delineates the Salar 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.

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^{92,93} of Chile and the Soquimich (SQM) mining company. Temperature value records are provided for the period from 1974 to 2019, relative humidity values are from 1987 to 2019 and precipitation volume data are from 1959 to 2019. The basic statistical indicators of the aforementioned meteorological variables can be seen in Table 1.

The precipitation samples for the isotopic analysis are compiled from previously published studies^{94–99} for the period from 2002 to 2021 for $\delta^{18}\text{O}$ and $\delta^2\text{H}$ 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. The main statistical indicators of $\delta^{18}\text{O}$ and $\delta^2\text{H}$ can be seen in Table 2. Refer to “Application on Salar de Atacama” for the input data file.

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 “Isocompy workflow”, are not needed.

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.

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 three 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.

The reader is referred to “Application on Salar de Atacama” 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.

Results and discussion

The first stage of statistical analysis shows that altitude and longitude are significant variables for temperature and relative humidity in all 3 months, while latitude is also significant in March (Table 3). This is consistent with the DICTUC¹⁰⁰ results. For precipitation, latitude and altitude are significant variables in the three summer

	Number of stations	Unit	Min	Max	Mean	Median	Std.dev
Temperature	28	°C	−4.3	22.6	12.8	13.2	5.5
Relative humidity	24	%	15	69.8	23.6	20.8	11.6
Precipitation	31	Mm	0	219	4.7	0	16.9

Table 1. Statistical indicators of temperature, relative humidity and precipitation in January, February and March for Salar de Atacama study area.

	Number of samples	Unit	Min	Max	Mean	Median	Std.dev
$\delta^{18}\text{O}$	52	‰ VSMOW	−15.2	−0.2	−8.1	−8.0	4.3
$\delta^2\text{H}$			−102.5	0.9	−53.3	−52.7	32.2

Table 2. Statistical indicators of $\delta^{18}\text{O}$ and $\delta^2\text{H}$ in January, February and March for Salar de Atacama study area.

```

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 10. 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.

```

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

months, as Houston and Harley¹⁰¹ 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^{86,102,103}.

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 3. Column $Ln(x+1)$ shows the models whose feature $Ln(x+1)$ values are used since they result in higher R-squared values.

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 3). 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. Figure 13 shows the observed-versus-estimated values of the three features in three months.

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

```

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 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).

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

Table 3. Results of the first-stage statistical analysis and models per month. The bold p values denote significant parameters (<0.05). ^aStandardized standard deviation of the cross-validation scores of the estimation models.

et al. (23 °C)^{102,104} 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)^{100,105}.

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.⁹⁹, 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^{84,86,106,107}, 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

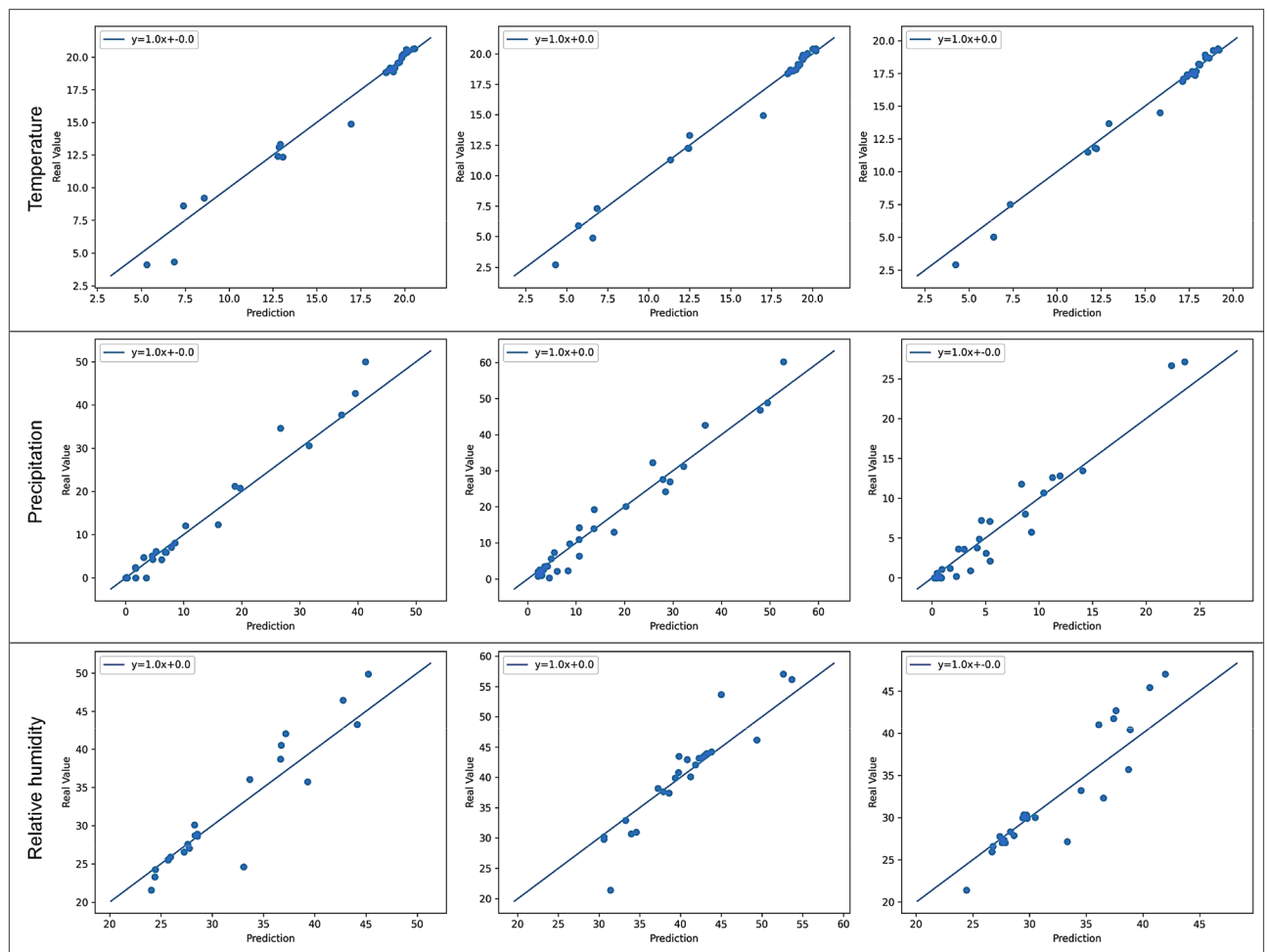


Figure 13. Plots of the estimated-versus-observed values generated by Isocompy for temperature, precipitation and relative humidity in January, February and March.

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^{90,100,105,108} 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)^{90,91,109}, as these studies considered all precipitation events during the year. In contrast, DGA⁸⁴ 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 $-0.19\text{‰}/100\text{ m}$ in summer (Fig. 14). This gradient is slightly lower than the others calculated in this region ($-0.34\text{‰}/100\text{ m}$ in Herrera et al. and $-0.26\text{‰}/100\text{ m}$ in Villablanca^{96,111}). 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^{98,112}.

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 4). 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 4. 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 4). 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.

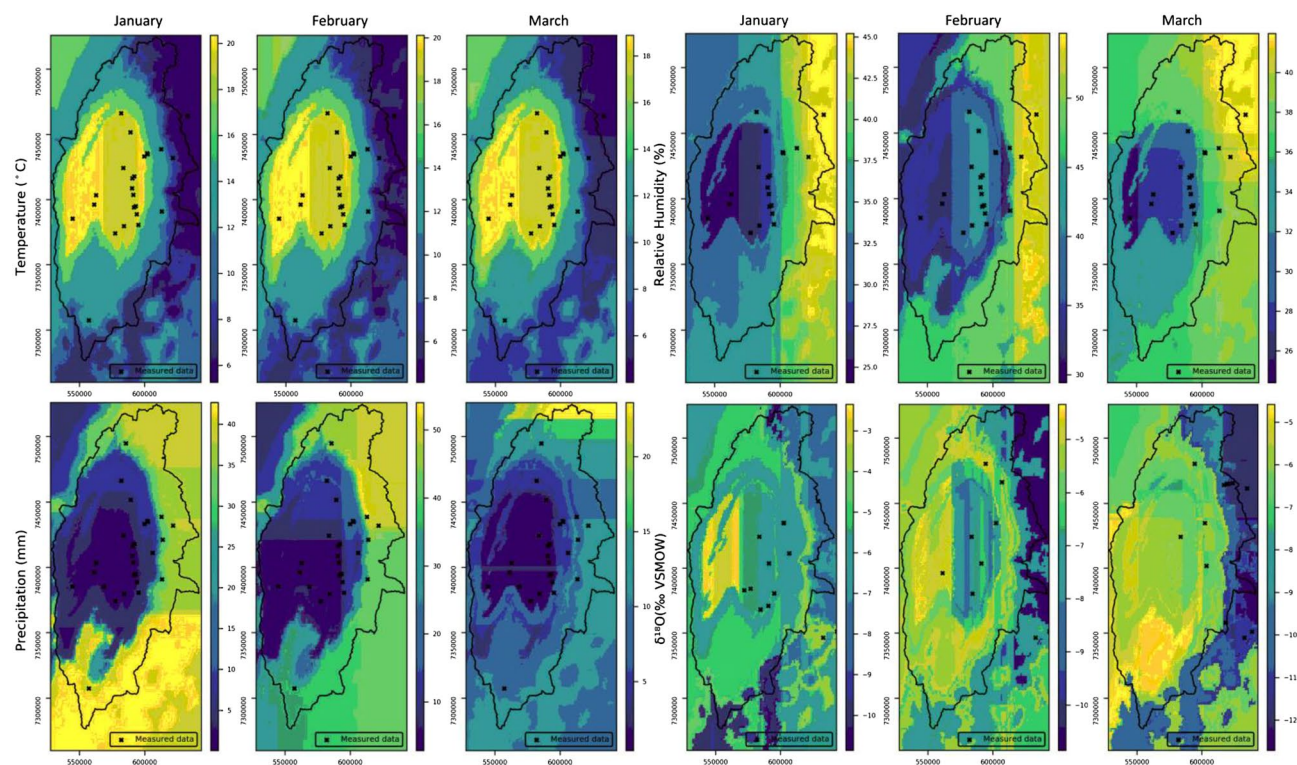


Figure 14. Maps of the temperature, precipitation, relative humidity and $\delta^{18}\text{O}$ values of precipitation estimated by Isocomp in January, February and March in the Salar de Atacama basin.

p-value	VIF_fin	VIF_init		Cor. lon	Cor. lat	Cor. alt	Cor. temp	Cor. prec	Cor. hum
–	–	8.6	Lon	1.00	–	–	–	–	–
0.78	1.1	2.0	Lat	–0.30	1.00	–	–	–	–
–	–	33.3	Alt	0.83	0.09	1.00	–	–	–
0.00	2.6	43.0	Temp	–0.87	–0.06	–0.98	1.00	–	–
0.66	2.3	2.6	Prec	0.60	–0.05	0.61	–0.60	1.00	–
0.04	3.7	4.5	Hum	0.78	–0.09	0.72	–0.77	0.75	1.00

Table 4. 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).

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^{110,113–117}. However, these differences are expected since the LMWL is calculated based on a different group of points in a larger area. Figure 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 Isocomp.

Conclusion

Isocomp 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 Isocomp 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

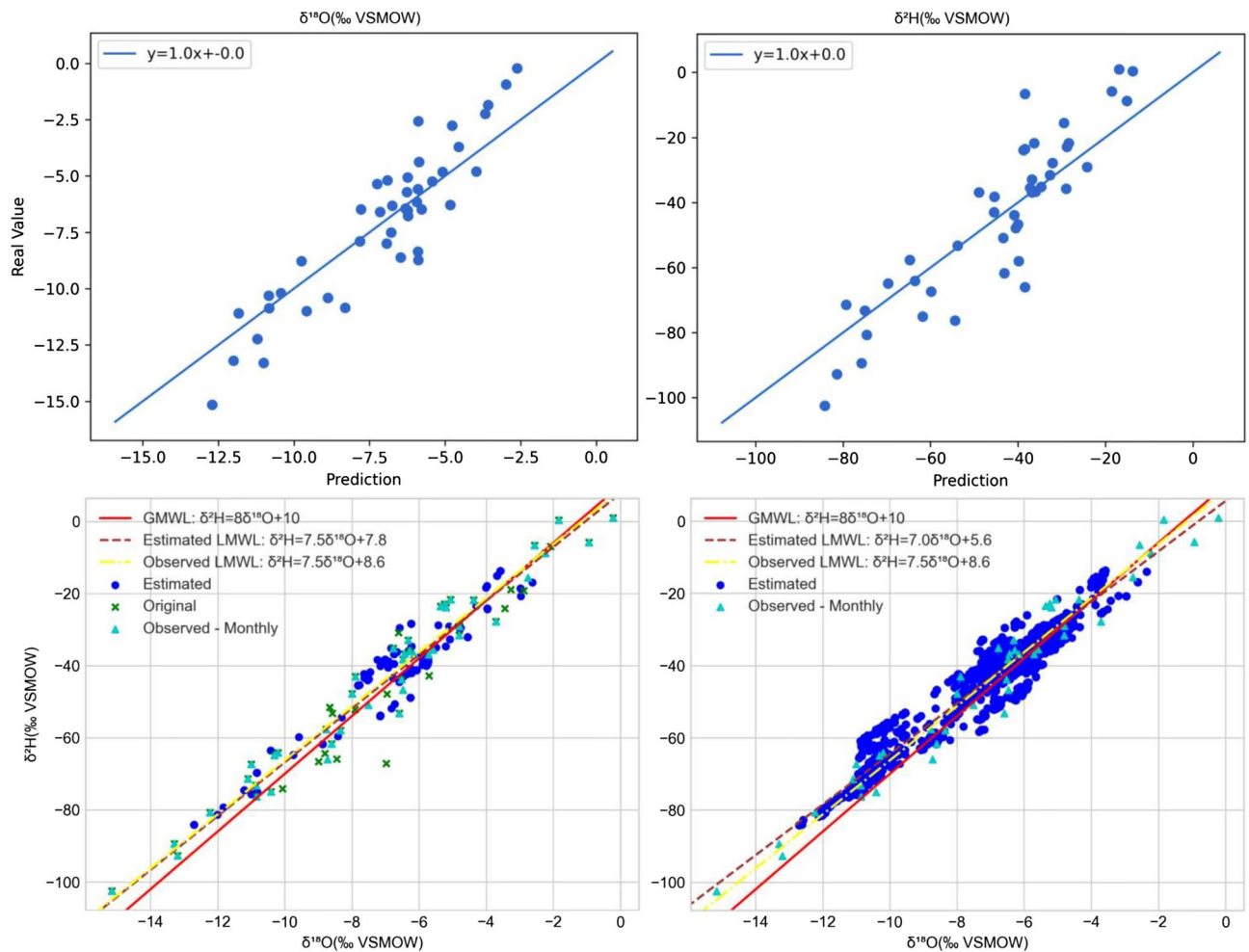


Figure 15. Top left and top right: estimations versus the measurements of $\delta^{18}\text{O}$ and $\delta^2\text{H}$, respectively. Bottom: plots of the estimated (circles) and observed (triangles) $\delta^{18}\text{O}$ versus $\delta^2\text{H}$ values of precipitation. The red line is the GMWL, the 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 “[Application on Salar de Atacama](#)” for the monthly meteoric line plots, residual plots and reports. All plots are generated by Isocompy.

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.

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

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>.

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.

Received: 7 September 2022; Accepted: 30 January 2023

Published online: 02 February 2023

References

- Aléon, J. *et al.* Determination of the initial hydrogen isotopic composition of the solar system. *Nat. Astron.* **2022**, 1–6. <https://doi.org/10.1038/s41550-021-01595-7> (2022).
- Custodio, E. & Llamas, M. R. *Hidrología Subterránea* (Omega, 1983).
- Custodio, E. & Jódar Bermúdez, J. *Recarga Natural a Los Acuíferos, Metodología y Soporte de la Isotopía del Agua*. (2019).
- 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).
- Merlivat, L. & Jouzel, J. Global climatic interpretation of the deuterium-oxygen 16 relationship for precipitation. *J. Geophys. Res.* **84**, 5029–5033 (1979).
- Araguás-Araguás, L., Froehlich, K. & Rozanski, K. Deuterium and oxygen-18 isotope composition of precipitation and atmospheric moisture. *Hydrol. Process.* **14**, 1341–1355 (2000).
- Jasechko, S. Global isotope hydrogeology—Review. *Rev. Geophys.* **57**, 835–965 (2019).
- Hurley, J. V. & Galewsky, J. A last-saturation diagnosis of subtropical water vapor response to global warming. *Geophys. Res. Lett.* **37**, 06702 (2010).
- 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).
- Risi, C., Bony, S. & Vimeux, F. Influence of convective processes on the isotopic composition ($\delta^{18}\text{O}$ and δD) of precipitation and water vapor in the tropics: 2. Physical interpretation of the amount effect. *J. Geophys. Res. Atmos.* **113**, 19305 (2008).
- 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).
- 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 Planet. Sci. Lett.* **307**, 47–58 (2011).
- 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).
- Craig, H. Isotopic variations in meteoric waters. *Science (80-)* **133**, 1702–1703 (1961).
- Feng, X., Faiia, A. M. & Posmentier, E. S. Seasonality of isotopes in precipitation: A global perspective. *J. Geophys. Res. Atmos.* **114**, 08116 (2009).
- Gat, J. R. Atmospheric water balance—The isotopic perspective. *Hydrol. Process.* **14**, 1357–1369 (2000).
- Gat, J. R. & Matsui, E. Atmospheric water balance in the Amazon Basin: An isotopic evapotranspiration model. *J. Geophys. Res.* **96**, 13179–13188 (1991).
- 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).
- Thomas, J. M. & Rose, T. P. Environmental isotopes in hydrogeology. *Environ. Geol.* **43**, 1 (2003).
- Cook, P. G. & Herczeg, A. L. *Environmental Tracers in Subsurface Hydrology. Environmental Tracers in Subsurface Hydrology*. <https://doi.org/10.1007/978-1-4615-4557-6> (Springer US, 2000).
- Coplen, T. *Stable Isotope Hydrology: Deuterium and Oxygen-18 in the Water Cycle*. *Eos, Transactions American Geophysical Union*. Vol. 63. (International Atomic Energy Agency, 1982).
- Kendall, C. & McDonnell, J. J. *Isotope Tracers in Catchment Hydrology*. <https://doi.org/10.1029/99eo00193> (Elsevier, 1998).
- Mook, W.G. *Environmental Isotopes in the Hydrological Cycle Volume 1.pdf. Technical Documents in Hydrology*. Vol. 1 (2000).
- 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).
- 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).
- 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).
- 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).
- Neale, R. B. *et al.* *Description of the NCAR Community Atmosphere Model (CAM 5.0). Ncar/Tn-464+Str 214* (2004).
- 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).

30. 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).
31. Kurita, N. *et al.* Intraseasonal isotopic variation associated with the Madden-Julian oscillation. *J. Geophys. Res. Atmos.* **116**, 24101 (2011).
32. 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).
33. 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).
34. 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).
35. 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).
36. 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).
37. 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).
38. 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).
39. Nguyen, L. D., Heidbüchel, L., Meyer, H., Merz, B. & Apel, H. What controls the stable isotope composition of precipitation in the Asian monsoon region? *Hydrol. Earth Syst. Sci. Discuss.* <https://doi.org/10.5194/hess-2017-164> (2017).
40. 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).
41. Rozanski, K., Sonntag, C. & Munnich, K. O. Factors controlling stable isotope composition of European precipitation. *Tellus* **34**, 142–150 (1982).
42. Liebmann, B. Description of a complete (interpolated) outgoing longwave radiation dataset. *Bull. Am. Meteorol. Soc.* **77**, 1275–1277 (1996).
43. Morales, M. S., Christie, D. A., Neukom, R., Rojas, F. & Villalba, R. Variabilidad hidrológica en el sur del Altiplano: Pasado, presente y futuro. *La Puna Argentina Nat. Cult.* **24**, 75–91 (2018).
44. 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).
45. Vuille, M. *et al.* Climate change and tropical Andean glaciers: Past, present and future. *Earth-Sci. Rev.* **89**, 79–96 (2008).
46. Stein, A. F. *et al.* NOAA's HYSPLIT atmospheric transport and dispersion modeling system. *Bull. Am. Meteorol. Soc.* **96**, 2059–2077 (2015).
47. 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).
48. 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).
49. Yoshimura, K., Kanamitsu, M., Noone, D. & Oki, T. Historical isotope simulation using reanalysis atmospheric data. *J. Geophys. Res. Atmos.* **113**, 19108 (2008).
50. Koh, K., Kim, S. J. & Boyd, S. A method for large-scale ℓ_1 -regularized logistic regression. *Proc. Natl. Conf. Artif. Intell.* **1**, 565–571 (2007).
51. Tipping, M. E. Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **1**, 211–244 (2001).
52. Efron, B. *et al.* Least angle regression. *ArXiv* <https://doi.org/10.1214/0090536040000006732.407-499> (2004).
53. MacKay, D. J. C. Bayesian nonlinear modeling for the prediction competition. *ASHRAE Trans.* **100**, 1053–1062 (1994).
54. Mallat, S. G. & Zhang, Z. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.* **41**, 3397–3415 (1993).
55. Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.* **10**, 61–74 (1999).
56. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006).
57. Hinton, G. E. *Connectionist Learning Procedures*. (1989).
58. Claesen, M. & De Moor, B. *Hyperparameter Search in Machine Learning*. (2015).
59. 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*. Vol. 2015. 1026–1034 (2015).
60. 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).
61. Chang, C. C. & Lin, C. J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**, 27 (2011).
62. 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).
63. Ploton, P. *et al.* Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nat. Commun.* **11**, 1–11 (2020).
64. Kraskov, A., Stögbauer, H. & Grassberger, P. Estimating mutual information. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.* **69**, 16 (2004).
65. Li, W. Mutual information functions versus correlation functions. *J. Stat. Phys.* **60**, 823–837 (1990).
66. Smith, R. A mutual information approach to calculating nonlinearity. *Statistics* **4**, 291–303 (2015).
67. *The Pandas Development Team*. pandas-dev/pandas: Pandas. 10.5281/zenodo.3509134 (2020).
68. Jordahl, K. *et al.* geopandas/geopandas: v0.10.2. 10.5281/ZENODO.5573592 (2021).
69. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
70. Haentjens, N. *pylr2 · PyPI*. <https://pypi.org/project/pylr2/>. Accessed 3 Mar 2022 (2018).
71. Seabold, S. & Perktold, J. statsmodels: Econometric and statistical modeling with python. in *9th Python in Science Conference* (2010).
72. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
73. 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).
74. Peterson, T. C., Vose, R., Schmoyer, R. & Razuvaev, V. Global historical climatology network (GHCN) quality control of monthly temperature data. *Int. J. Climatol.* **18**, 1169–1179 (1998).
75. Somaya, H. & Tomader, M. *Tuning the Hyperparameters for Supervised Machine Learning Classification, to Optimize Detection of IoT Botnet*. 1–6. <https://doi.org/10.1109/ISIVC54825.2022.9800742> (2022).
76. Gillies, S. *et al.* *Shapely: Manipulation and Analysis of Geometric Objects*. (2007).
77. Bokeh Development Team. *Bokeh: Python Library for Interactive Visualization*. (2018).
78. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

79. Friedman, J., Bohonak, A. J. & Levine, R. A. When are two pieces better than one: Fitting and testing OLS and RMA regressions. *Environmetrics* **24**, 306–316 (2013).
80. Chen, F. *et al.* Local meteoric water lines in a semi-arid setting of northwest China using multiple methods. *Water* **13**, 2380 (2021).
81. 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).
82. 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. <https://doi.org/10.25080/majora-ebaa42b7-00d> (2011).
83. Amfios21. *Estudio de Modelos Hidrogeológicos Conceptuales Integrados, para los Salares de Atacama, Maricunga y Pedernales. Etapa III. Informe Final. Modelo Hidrogeológico Consolidado Cuenca Salar de Atacama.* (2018).
84. DGA. *Análisis de la Oferta Hídrica del Salar de Atacama. Sdt No. 339* (2013).
85. DGA. *Evaporación desde salares: Metodología para Evaluar Los Recursos Hídricos Renovables. Aplicación en las Regiones I y II. Vol. 1.* (Revista de la Sociedad Chilena de Ingeniería Hidráulica, 1986).
86. Hess, R. A. Simplified approach for modelling pilot pursuit control behaviour in multi-loop flight control tasks. in *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*. Vol. 220 (2006).
87. 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).
88. 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).
89. 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).
90. IDAEA-CSIC. *Cuarta Actualización del Modelo Hidrogeológico del Salar de Atacama.* SNIFA. <https://snifa.sma.gob.cl> (2017).
91. 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).
92. DGA. *Servicios Hidrometeorológicos.* <https://www.dga.cl/servicioshidrometeorologicos/Paginas/default.aspx> (2020).
93. Centre for Climate and Resilience Research. *Datos de Precipitación, Datos de Temperaturas.* <https://www.cr2.cl/datos-de-precipitacion/> (2018).
94. Cortecci, G. *et al.* New chemical and original isotopic data on waters from El Tatio geothermal field, northern Chile. *Geochem. J.* **39**, 547–571 (2005).
95. 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.* <https://doi.org/10.1079/BJN20041276> (2017).
96. 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).
97. 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, Thesis for Degree of Master of Sciences Mention in Geology, 2016).
98. 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).
99. 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).
100. 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).
101. 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).
102. 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).
103. Vázquez, Enric & Ayora, C. *Cuarta Actualización del Modelo Hidrogeológico del Salar de Atacama.* (2017).
104. 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).
105. MOP-DGA. *Balace Hídrico de América del Sur.* (1988).
106. 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).
107. 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).
108. SGA. *Estudio Hidrogeológico y Modelo Numérico sector sur del Salar de Atacama.* (2015).
109. 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).
110. Geol, X. I. I. C., Santiago, C., Geol, C., Ambiente, M. & Cient, S. Estudio de la relación isotópica $\delta^{18}\text{O}/\delta^2\text{H}$ de los manantiales en el sector de las nacientes del Loa, Región de Antofagasta. in *XII Congress Geológico Chile*. 16–19 (2009).
111. Villablanca, D. Estudio de la relación isotópica $\delta^{18}\text{O}/\delta^2\text{H}$ de los manantiales en el sector de las nacientes del Loa, Región de Antofagasta. in *XII Congress Geológico Chile*. 22–26 (2009).
112. 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).
113. Fritz, P., Suzuki, O., Silva, C. & Salati, E. Isotope hydrology of groundwaters in the Pampa del Tamarugal. *Chile J. Hydrol.* **53**, 161–184 (1981).
114. Aravena, R. *et al.* Isotopic composition and origin of the precipitation in Northern Chile. *Appl. Geochem.* **14**, 411–422 (1999).
115. 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'Inst. Français d'Études Andin.* **27**, 367–384 (1998).
116. Chaffaut, I. *Précipitations 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).
117. 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**, 4 (2019).

Acknowledgements

The authors acknowledge Carlos Ayora and anonymous reviewers that helped us to improve this article. This study was supported by the “Agencia Estatal de Investigación” from the Spanish Ministry of Science and Innovation and the IDAEA-CSIC, a Centre of Excellence Severo Ochoa (CEX2018-000794-S). R. Criollo gratefully

acknowledges the financial support from the Balearic Island Government through the Margalida Comas post-doctoral fellowship programme (PD/036/2020).

Author contributions

A.H., S.V., E.V.-S. and R.C. devised the conceptual idea. A.H. developed the algorithm and the code. A.H. and S.V. contributed to the validation of the code, example application and interpretation of the results. A.H. took the lead in writing the manuscript. S.V., E.V., R.C. and M.C. provided critical feedback and helped shape the manuscript. This research article is part of the PhD thesis of A.H. at the Geology Program, UAB.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.H.

Reprints and permissions information is available at www.nature.com/reprints.

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023