Technical Report – Methods: Automated Discovery of Functional Relationships in Earth Systems Data

R. Reinecke[1,2], F. Pianosi[3,4], and T. Wagener[1]

[1]Institute of Environmental Science and Geography, University of Potsdam, Potsdam, Germany

[2]Institute of Geography Johannes Gutenberg-University Mainz, Mainz, Germany

[3]Department of Civil Engineering, University of Bristol, Bristol, UK

[4]Cabot Institute, University of Bristol, Bristol, UK

Corresponding author: Robert Reinecke (reinecke@uni-mainz.de)

Key Points:

- Functional relationships capture how variables co-vary across spatial or temporal domains.
- Here we present a new method for the automated diScovery Of fuNctionaAl Relationships (SONAR).
- We test SONAR on model-derived datasets to identify functional relationships of groundwater recharge simulations from global hydrological models with possible drivers.
- We compare SONAR to two established methods, CART (Classification and Regression Trees) and CIT (Conditional Inference Trees), and find that SONAR produces smaller trees and is more robust.

Abstract

Functional relationships capture how variables co-vary across specific spatial or temporal domains. However, these relationships often take complex forms beyond linear, and they may only hold for sub-sets of the domain. More problematically, it is often a priori unknown how such sub-domains are defined. Here we present a new method called SONAR (diScovery Of fuNctionaAl Relationships) that enables the automated discovery of functional relationships in large datasets. SONAR operates on existing unstructured data and is designed to be an explorative tool for large datasets where manual search for functional relationships would be impossible. We test the method on groundwater recharge outputs of several global hydrological models to explore its usefulness and limitations. Further, we compare SONAR to the established CART (Classification and Regression Trees) and CIT (Conditional Inference Trees) methods. SONAR results in smaller trees with functional relationships in the leaf nodes instead of specific classes or numbers. SONAR provides a robust and automated method for the exploration of functional relationships.

Plain Language Summary

Vastly expanding datasets have the potential for incredible advancements in our understanding of how different variables co-vary within Earth system dynamics. However, we lack adequate tools to identify new relationships within such complex and high-dimensional datasets. Here we developed a new method called SONAR that can automatically find relationships in large datasets. We test the method on global simulations of groundwater recharge and find that it produces smaller and more robust structured representations than existing methods. SONAR is an exploratory tool that can help researchers discover relationships in complex datasets in the Earth sciences and beyond.

1 Introduction

Earth system science relies on understanding functional relationships, which can be defined as the co-variation of variables across space or and time that underpins our theoretical knowledge of how the Earth works (Gnann et al., 2023a; L'vovich, 1979). For example, we find that groundwater recharge across water limited domains co-varies with available precipitation (MacDonald et al., 2021), or that changes in the co-variation of precipitation and runoff can reflect system changes in response to drought (Peterson et al., 2021). To understand and anticipate the evolving Earth system (Denissen et al., 2022), we require a quantitative understanding of this co-variation. Not only is an understanding of such relationships important for our scientific understanding, it also allows us to build adequate models and evaluate their consistency with the Earth system dynamics we observe (Eker et al., 2018; Koster & Milly, 1997; Reichstein et al., 2019; Wagener et al., 2022). If finding functional relationships offers such a high reward, how do we find them beyond manually looking for them – given that we can rarely identify them through planned experiments at our scales of interest?

The dramatic increase in the size of datasets describing the structure and dynamics of the Earth system offers huge opportunities for finding new relationships - if we have the tools to identify them in vast and complex data. We have increasingly large satellite datasets; for example, the new SWOT mission will send more than 1TB per day back to Earth, and the NASA Earth data repository is estimated to grow to over 245 PB by 2025 (NASA, 2021). This does not even include model outputs which add even more to the pile of data we have (e.g. Hoch et al. (2023)).

It will not be feasible to manually search through such datasets for functional relationships – unless one makes very strong and thus limiting a priori assumptions about what we expect to find. On the other hand, we struggle with imbalanced data, i.e. we often have unequal distributions of relevant classes within the data (Bradter et al., 2022; Chawla et al., 2002; Kaur et al., 2020), with human interference (Krabbenhoft et al., 2022), and with epistemic uncertainty (Beven et al., 2018; Beven & Cloke, 2012). For example, Krabbenhoft et al. (2022) show that global streamflow observations are significantly imbalanced and globally organized more by national GDP than by hydrological considerations, thus providing limited information in dry regions.

Earth systems datasets are a mixture of organized sampling (e.g. some remotely sensed observations) and those that are not sampled in a strategic manner, but are rather samples of opportunity (e.g. groundwater recharge estimates), thus requiring analysis methods that can work with all samples. Methods that can work with generic input-output datasets have been called sampling-free or data-agnostic methods (Pianosi & Wagener, 2018; Sheikholeslami & Razavi, 2020). Further, if methods require no manual parameter tuning, we call them parameter-free (Saltelli et al., 2021). This is another advantageous feature of a method given that parameter tuning can be different if very heterogenous and imbalanced datasets are studied. Both properties would be beneficial for the automated exploration of functional relationships in Earth system data.

Earth system processes are driven by different factors across space and time scales (Pattee, 1972), vary along gradients (Lesk et al., 2021), and exhibit thresholds (Zehe & Sivapalan, 2009). Thus, an automated method should also be able to identify and represent relationships in a hierarchical manner to represent the diversity in subdomains of the data. In the past, tree-like algorithms such as CART (Classification and Regression Trees) (Breiman et al., 2017) and CIT (Conditional Inference Trees) (Hothorn et al., 2006) and other similar implementations (Loh, 2014) have been used to find hierarchical structure in Earth system data (e.g., Messager et al. (2021), Almeida et al. (2017)). While these algorithms have initially been built for classification and regression, they also provide information about dominant controls. In fact, the point at which the data are split into subtrees reveals the underlying structure of the data and the dominant controls that separate sub-domains. However, these data-based strategies can show limited robustness and can provide splits at non-physical boundaries rendering their interpretation difficult (Sarailidis et al., 2023).

Addressing the robustness problem, ensemble methods such as random forest (Breiman, 2001) can identify dominant controls through factor importance (Antoniadis et al., 2021), while others have used multivariate adaptive regression splines (MARS) (Friedman, 1991) to find more complex relationships (e.g., Conoscenti et al. (2015)). However, such approaches can be difficult to interpret or even visualize. While visual inspection remains powerful in identifying complex variable interactions – especially if we do not know what kind of interaction we might expect (Puy et al., 2022; Wagener & Kollat, 2007). Similarly, machine learning has led to approaches that learn functional relationships (Shrestha et al., 2009), and explainable AI strategies are advancing rapidly (Jiang et al., 2022).

Here we present an automated method for the di**S**covery **O**f fu**N**ctiona**A**l **R**elationships (SONAR) that combines data agnosticism, interpretability, and the identification of hierarchical controls, in a parameter-free algorithm. What distinguishes SONAR from other existing methods is that the automatic search yields a tree that separates the search domain in a hierarchical manner and uncovers possible functional relationships. To our knowledge, no method exists that can automatically separate data in a hierarchical manner to show functional relationships. SONAR is tested here on a large groundwater recharge dataset from eight global hydrological models.

Groundwater recharge is an example of a hydrological process (see supplement for definition) which remains highly uncertain on the global scale as hydrological models disagree largely in the functional relationships they produce (Berghuijs et al., 2022; Reinecke et al., 2021; West et al., 2023). It is unclear why exactly the models disagree and how it relates to differences in assumptions made about how hydrologic systems work. However, one can clearly trace patterns of different recharge behavior for different climatic zones across the globe (Fig. S1). Here we test whether SONAR can be used to analyze synthetic (noise-free) datasets produced by hydrological models and identify different functional relationships in different sub-domains (e.g. climatic regions); and how its results compare with established strategies.

2 Materials and Methods

2.1 Automated discovery of functional relationships

SONAR works similarly to other tree-based approaches such as CART (Breiman et al., 2017). However, SONAR is not built to solve a classification or a regression problem but to find functional relationships while making no prior assumption about the type of relationship beyond a choice of correlation metric (that can be varied; in the following we use the spearman rank correlation). The algorithm works as follows (Fig. 1). It searches recursively for the best possible split within the dataset. On each split SONAR determines which binary separation of an explanatory variable (e.g., amount of precipitation above or below a certain threshold) would increase the correlation between an explanatory variable (e.g., aridity index, or precipitation amount again) and the variable under investigation (e.g., groundwater recharge). SONAR searches for possible splits based on equally sized bins to reduce the search space into manageable pieces. However, the correlations are always calculated on the original data and not the bins. SONAR tests all possible splits based on different subsets of the bins (Fig. 1) from small to large values of the explanatory variables (for description of alternatives see Supplement). SONAR can also handle categorical variables, in which case the split is based on whether the data belong to a certain category or not. With each split SONAR searches for an increase in correlation. SONAR produces binary trees and for each split at least one side (the left or right subtree) needs to increase in correlation otherwise the algorithm stops (Fig. 1). Requiring an increase for both sides would yield a less robust algorithm given that we want to distinguish sub-domains in which functional relationships exists from those where this is not the case. To ensure that SONAR does not select very small subspaces a split requires each subspace to have at least 500 data points or 5% of the data of the parent node – depending on the dataset used. This value can be changed and limits the parameter-free property of the approach. Importantly, each leaf node ends up containing a relationship and not only a particular class (compared to classification trees) or value (compared to regression trees). Each leaf thus contains a subset of the original data points for the particular subdomain. SONAR then derives a functional relationship in the following way: the data in each leaf node are divided into 10 equally-sized bins and a line is added that connects the medians across the bins to describe the functional relationship.

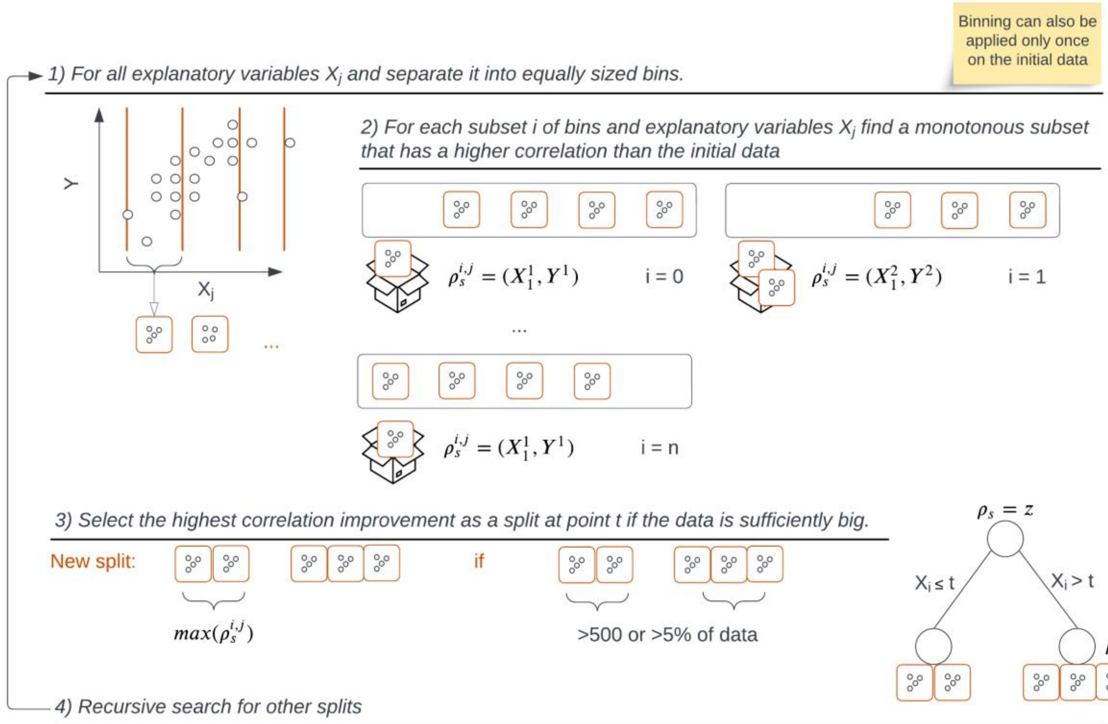

Figure 1. Visual representation of the SONAR algorithm and its major workflow components. Y denotes the variable we are searching dominant controls for in the set of explanatory variables Xj. ps is the Spearman Rank correlation and z the highest ps of the node above a split (this can also be the root node).

2.2 Approaches related to our method: CART and CIT

We compare our approach to two existing methods: CIT (Conditional Inference Trees) (Hothorn et al., 2006) and CART (Classification and Regression Trees) (Breiman et al., 2017). We selected these two methods because CART is well established and widely used, while CIT is conceptually closest to our method as it searches for correlations as well, though without the explicit search for functional relationships. Ensemble methods such as Random Forest (Breiman, 2001) are more complex realizations of the single tree methods used here but have the above discussed problems of interpretability, hence we do not include them here. MARS (Multivariate Adaptive Regression Splines) (Friedman, 1991) and other regression methods cannot separate domains in a hierarchical manner.

Using a greedy approach (A selection of the best possible option at a current state of the algorithm, thus possibly missing a global optimum), CART searches for an optimal binary split of a dataset that optimizes an error function such as the Gini index or an entropy measurement. CART trees tend to overfit and thus must be pruned for most datasets (Esposito et al., 1997). CIT is similar to CART as it constructs a binary tree and can produce regressions and classifications. However, to decide on a split CIT tests for a maximum linear independence between covariates and response variables. CIT stops if the null hypothesis H0 of variable’s independence cannot be rejected. It selects a subset of the covariate with the highest conditional expectation using a linear two-sample test. CIT can be computationally expensive and was in the past used, e.g., to determine the role of global change in soil functions (Rillig et al., 2019). It was, however, criticized due to its limited ability for detecting non-linear effects (Wright et al., 2017).

In both CART and CIT trees, dominant controls are indicated by variables close to the tree's root node. The earlier a variable is used for a split the more a separation improves the classification or regression fit. Splits in SONAR provide a similar indication, however, controls also appear in the leaf nodes. The controls selected in the leaf nodes may be equal to the ones used for a split or be different.

2.3 Experimental setup

2.3.1 Groundwater recharge data and explanatory variables

We use groundwater recharge (see S1) as an example process to test the algorithms. Groundwater recharge is poorly understood globally and available data are rather imbalanced (Gnann et al., 2023a). For these reasons we use data produced by model simulation, rather than observations. We also use a long-term estimate of recharge given that this is most likely related to climatic factors which we consider here. Our dataset consists of simulated 30-year annual averages of groundwater recharge on a 0.5° spatial resolution from an ensemble of eight global hydrological models (Table S1) (Best et al., 2011; Burek et al., 2020; Gnann et al., 2023a; Hanasaki et al., 2018; Müller Schmied et al., 2021; Schaphoff et al., 2018; Sutanudjaja et al., 2018; Swenson & Lawrence, 2015; Takata et al., 2003). We investigate functional relationships within the data to showcase differences between the algorithms. There is no intention here to evaluate the specific model implementations or performances. For the classification task of CART, we separate annual groundwater recharge amounts into four classes: very low (0-10 mm/yr), low (10-100 mm/yr), medium (100-500 mm/yr), and high (>500 mm/yr). Using different separation categories does not change the general conclusions regarding the algorithms but influences the specific CART trees (see Fig. S13). All models are driven with the same forcing input (Table S2). Recharge simulations and forcing data are based on the simulation protocol of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP) (Warszawski et al., 2014).

In addition, we use a set of explanatory variables that we assume to be potentially relevant in determining recharge in the eight models (Table S2 and Fig. S5-S9). We use long-term mean precipitation (P), long-term mean potential evapotranspiration (PET), an aridity index (AI) defined by PET/P, long-term mean temperature (T), an indicator of cold days per year (DB), and a land cover data set GlobCover which is closest to the information used in the models (ESA, 2010). In contrast to common forcing, the hydrological models used consider very different geological information which is therefore hard to consider here.

Traditionally machine learning methods are evaluated with established datasets like Iris (Unwin & Kleinman, 2021) or Forest cover type (Jock Blackard, 1998), however they are either too small to be used with SONAR or are built specifically for a classification problem which cannot test the usefulness of approach.

2.4 Evaluation criteria of method attributes

2.4.1 Comparison between SONAR, CART and CIT

The three methods include different information in their leaf nodes and make very different split decisions (see Section 2.2). To allow a general comparison, we compare the trees visually in their pathways to derive at certain recharge classes (see 2.3.1). We focus on the dominant controls (how far up in the tree explanatory variables are mentioned; see also 2.2), their thresholds (split decisions), and the pathways that lead to certain value ranges. For the widely used Iris dataset (Unwin & Kleinman, 2021) and a simple CART tree this path representation shows that petal width is a dominant control (Fig. S14)

Since no other existing method represents functional relationships in their leaf nodes we use the derived functional line of SONAR (see 2.1) to calculate ranges of values within the node (i.e., the range of possible Y for a given range of X) that can be compared to the regression and class ranges of CART and CIT.

2.4.2 Robustness of SONAR

To test how SONAR reacts to data limitations we create a robustness test. A possible real-world reason for this absence of data could be a sampling bias (e.g. Krabbenhoft et al. (2022)). Each experiment removes a certain percentage of data from the original dataset at random. The less a tree representation changes the more robust the algorithm is. This does not address the correctness of the tree. We measure the robustness by utilizing the TED (tree-edit-distance) (Pawlik & Augsten, 2015) defined as the minimum-cost sequence of node edit operations (delete, insert, rename) that transform one tree into another. We use TED only to compare trees derived within a method and not for cross-method comparison. In 100 independent experiments, 1 is the baseline experiment with all the available data, we randomly remove X% of the initial data and compare the resulting tree to the baseline experiment. A method is more robust to random removal of data if the TED remains small between the baseline and the 99 other experiments. As a reference we compare the robustness of SONAR with the widely used CART method.

3 Results

3.1 Automatic detection of relationships in sub-domains using SONAR

Testing SONAR on groundwater recharge datasets from eight global hydrological models yields eight different trees, two of which are shown in Fig. 2. We show models WaterGAP (Müller Schmied et al., 2021) and LPJML (Schaphoff et al., 2018) (see also Table S1) as examples, while all other models can be found in supplement S5. All resulting trees are rather shallow with only one to four splits. This is a characteristic of SONAR that is amplified by the minimum number of points requirement (see 2.1; without it the trees grow only marginally bigger, see supplement S5).

SONAR finds highly correlated subsets of the data in its leafs with Spearman rank correlations $p_s > 0.9$ (up to 0.95 for model (a) in Fig. 2a). Separation into different subspaces of the explanatory variables, by temperature in Fig. 2a and by aridity index in Fig. 2b, together with the different functional relationships in the leaf nodes, suggests that the global models WaterGAP and LPJML differ in the way they represent groundwater recharge processes.

In Fig. 2a, the dominant control for the tree is the aridity index in all leaves; for the tree in Fig. 2b, it is precipitation. The fact that the same control appears in all leaves within a tree is specific to these two trees, and different controls will be found across other datasets. Compared to the initial correlation of 0.89 and 0.77 at the root node (both to precipitation), the correlation increases for some subdomains but decreases for others. (SONAR only requires an increase in one subdomain on a split, see 2.1). In our case study, the number of points in the highly correlated domains is always much smaller than those in the less correlated domains and also shows higher uncertainty in the functional relationships found (Fig. 2).

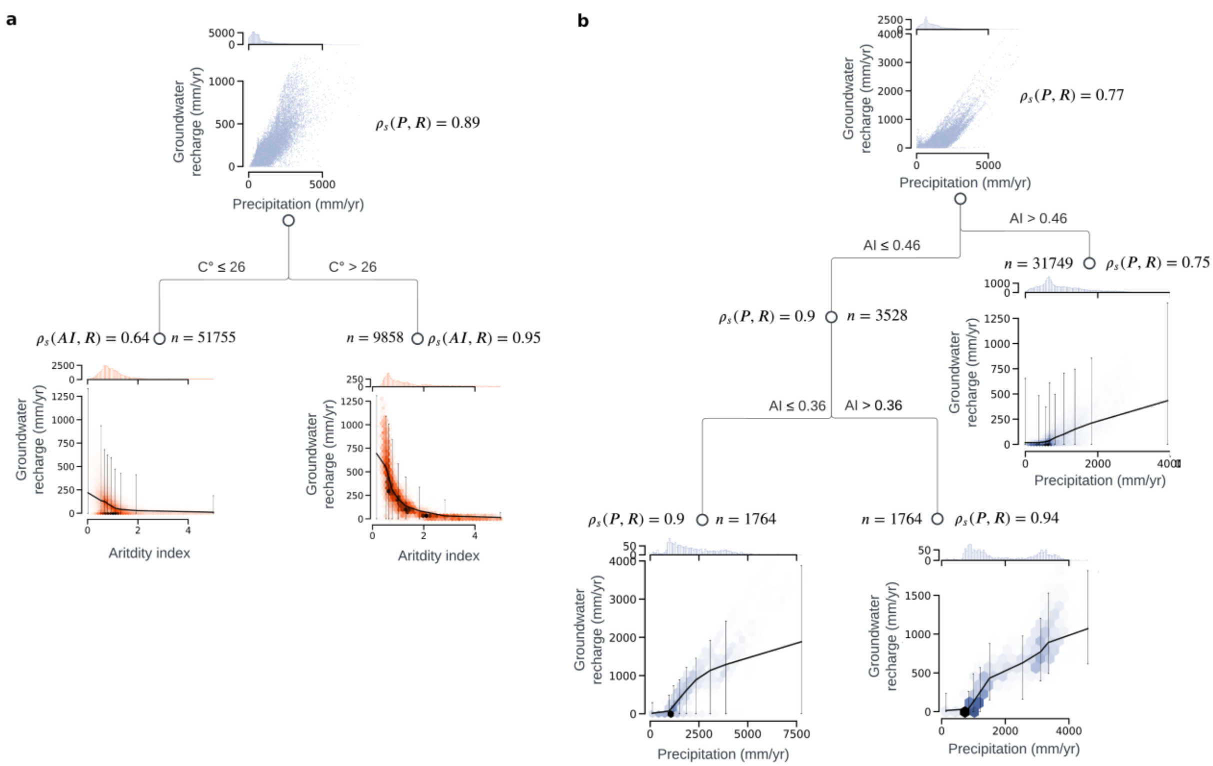

Figure 2. SONAR tree of models WaterGAP (a) and LPJML (b). n is the number of points at each node, p_s the spearman correlation, the black line is the functional relationship, error bars indicate the min. and max. value in each bin (here 10 quantiles). The color provides an indication of the point density of the underlying data as a visual aid (lines and error bars are calculated based on the underlying scatter of the original data). The darker the color the more points are inside this area. The root shows the relationship between Precipitation (P) and Recharge (R) because this shows the highest initial correlation in the data without splits.

To ensure that SONAR finds reasonable relationships we tested it with the same explanatory variables and (1) randomly generated recharge, (2) recharge generated based on linear relations to precipitation that differ for different domains, and (3) recharge generated based on PET (see supplement). Using these examples, we show that SONAR does not produce any tree from randomly generated data and is able to identify the artificial relationships for precipitation and PET (see supplemental S7).

3.2 SONAR differs from CART and CIT in regression and classification paths

SONAR searches for functional relationships instead of classifications or regressions; nevertheless, the meanings of the trees are similar enough to CART and CIT to compare the interpretations and conclusions drawn. In Fig. 3, we represent sub-trees to enable such a comparison (for a full explanation of the chosen visualization, see supplemental material), including the results shown in Fig. 2a. For each tree, Fig. 3 only shows the part of the tree that describes controlling variables on recharge values smaller than 100 mm/yr as an example (see supplement Fig. S15, S16 for the complete trees). The visualization shows each path that leads to a recharge value below or equal to 100 mm/yr, from the first split at the root node (left) to the leaf node (right). A different box indicates a split, while the value and color inside the box indicate at which point and through which variable the data was split. If a box is bigger, there are more pathways and leaves following this split in the tree. The leaf shows only a single class for classification trees (CART), values below the chosen threshold for regressions (CIT), and a range of values within a functional relationship that produces values below the threshold (SONAR).

Equal to Fig. 2a the SONAR tree shows only one split at 26 C° in comparison to CART and CIT, which show more possible pathways to low recharge values. All three approaches show different dominant controls and pathways to low recharge values. The encoding of how low recharge values are reproduced is much more complex in CART and CIT (multiple splits and different variables that control them) and very short in SONAR. The CART tree suggests that precipitation is the dominant control (as it shows up earlier in the tree) and that the aridity index gets more important in certain subdomains. On the other hand, CIT also uses precipitation as the first split but other explanatory variables for splitting the data further. Overall all three methods differ substantially in their understanding of the data.

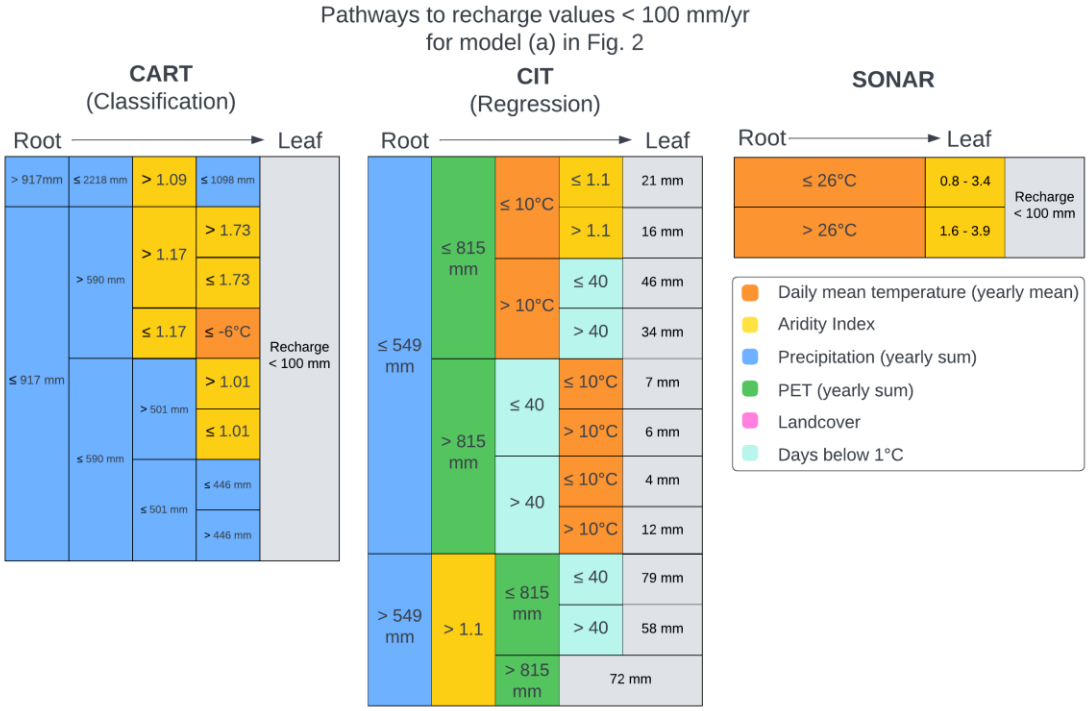

Figure 3. Visual representation of tree pathways (see supplement S4 for an extended explanation and simple example of this visualization method) only for low recharge values of three different approaches. The SONAR sub-plot shows part of Fig. 2a. For CART and CIT only, the part of the tree that leads to low values is shown. Gray boxes indicate the values or classes – for CIT and CART they are also the leaf nodes. All three trees were trained on the same model data and explanatory variables. The CART and CIT tree were pruned to a depth of 4.

3.3 SONAR is robust to variations in the input dataset

To test the robustness (see 2.4.2) of SONAR we removed a percentage of the original data and compared it with a baseline experiment. To provide a frame of reference we first conducted the experiment with the established CART algorithm (Fig. 4a). With an increased loss of information, the resulting CART trees become increasingly different (higher TED) from the baseline experiment which includes all data. Notably the mean difference between the models is relatively stable throughout. In comparison, SONAR is relatively robust as the TED with 10% loss is 1 magnitude smaller than with CART. Even with 50% of data loss SONAR only reaches a maximum TED of 5, for some models the tree does not change at all. Importantly, the small TED is likely highly impacted by the total size of the tree. SONAR leads to smaller trees to begin with.

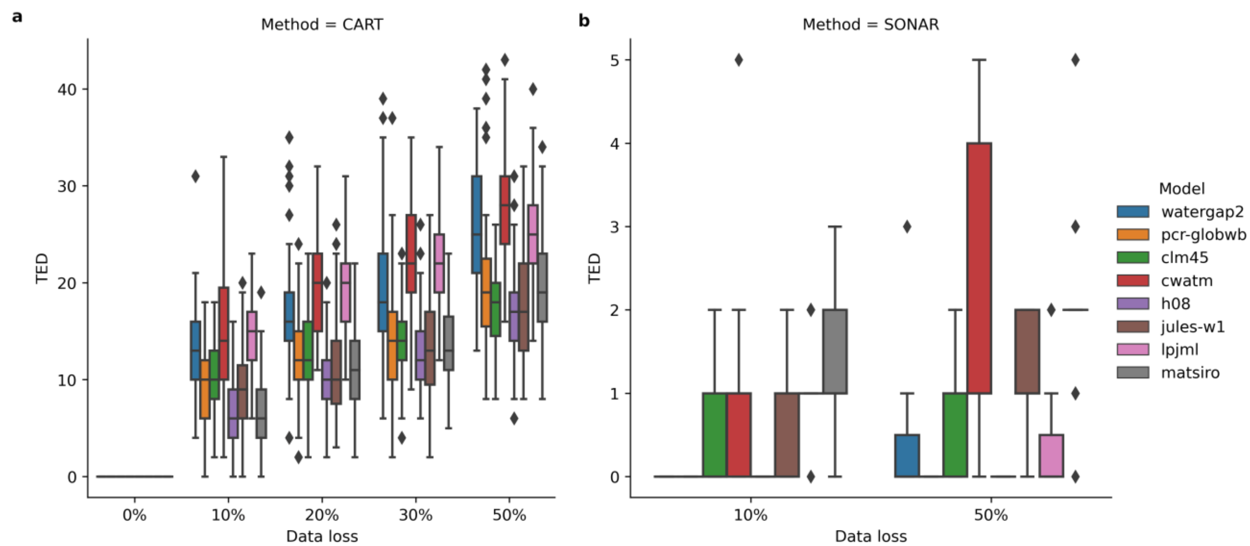

Figure 4. Robustness test of CART (a) and SONAR (b). Bars show the distribution of TED over the 99 independent random experiments as an indicator for robustness (small values equal a smaller change from the original tree). If the there is no bar shown the TED is 0 and all trees are equal for that model.

4 Discussion and method limitations

The application of SONAR to simulated groundwater recharge of global hydrological models shows differences between models and overall precipitation as a strong control of recharge. Both of these findings alight with recent analysis of this data (Gnann et al., 2023a; West et al., 2023). Importantly, SONAR also reveals that precipitation is not always the strongest explanation for recharge variability (Fig. 1a shows aridity as functional control of recharge) and that relationships between precipitation and recharge may differ across data subsets (e.g., divided by climate as in Fig. 1b). As recharge is a complex process which is not only controlled by available water but also by e.g. soil conditions and energy availability, one should expect different functional relationships in different domains (e.g. climatic regions). Model developers could use the identified relationships to evaluate whether their model represents a functional relationship that is similar to our hydrologic understanding and data of a specific region.

The analysis reveals that SONAR produces very robust small trees but also differs largely in the path found towards small recharge values from very established algorithms. Importantly, because SONAR is so different from other algorithms (a search for functional relationships instead of regression or classification), a comparative analysis can only provide limited insights into whether it is more useful than established algorithms. SONAR results might allow for an easier discussion of their hydrological meaning compared to e.g. CART due to the smaller trees and relationships instead of discrete classes in its leaves.

We did not investigate observational data at this stage and we did not extend the analysis to the temporal domain, but there would not be any fundamental difference in workflow. An important aspect that needs further consideration is the role of epistemic uncertainty when applying SONAR to observational data. However, SONAR does not produce any tree from randomly generated data (supplement S7) and is able to identify the artificially introduced relationships of precipitation and PET (supplement S7). Wider analysis to other datasets will be required to understand what relationships can be identified by SONAR.

The current implementation of SONAR has multiple limitations as we made specific methodological choices. Foremost, we could have used another correlation metric (Lee Rodgers & Nicewander, 1988), e.g., Pearson (Barber et al., 2020) instead of Spearman rank correlation. Also, metrics that consider a degree of regression fit would be possible. Our current choices are meant to require minimum assumptions. Furthermore, we chose to introduce a constraint on the amount of points at which a split is carried out, to prevent the algorithm from creating very small datasets in which the correlation calculation can become meaningless (see also S6). Selection of meaningful subset of data is an active field of research thus other approaches in separating the data at splits in SONAR could be considered (García-Pedrajas, 2011). And finally, the selection of explanatory variables has an impact on the results for any type of empirical algorithm like the one we present here, e.g. because variables like precipitation and aridity index are slightly correlated (Fig. S12).

5 Conclusions

SONAR describes a new and simple approach to identify functional relationships in complex datasets, thus giving effective insight into dominant controls within subdomains. The key advantage of SONAR is the automatic, non-parametrized, representation of functional relationships of hierarchical domains. It is specifically not built for classification or regression tasks, but to find possible relationships in large datasets. A comparison to other tree approaches shows that SONAR produces trees that are shorter and thus likely easier to interpret. Furthermore, SONAR is very robust and does not require any parameter tuning to work on a specific dataset.

Without any prior knowledge, SONAR enables researchers to explore vast datasets of model simulations and observations to automatically discover exciting new functional relationships. Especially in the field of hydrology, where controls differ largely across temporal and spatial domains, we demonstrated that this new method can yield interesting new insights. Eventually SONAR could also be used for model evaluation by enabling the comparison of functional relationships identified in the data to those identified in model simulations.

Acknowledgments

We thank Sebastian Gann for the discussions on functional relationships and valuable comments. RR and TW were funded by the Alexander von Humboldt Foundation in the framework of the Alexander von Humboldt Professorship endowed by the German Federal Ministry of Education and Research. FP was partially funded by the Engineering and Physical Sciences Research Council (EPSRC) "Living with Environmental Uncertainty" Fellowship (EP/R007330/1). RR designed and conducted the experiments and wrote the initial draft. TW had the initial idea. RR, TW and FP designed the method jointly. All authors contributed equally to the final manuscript.

Open Research

The original non-aggregated model data is available from isimip.org. The aggregated data is available at Gnann et al. (2023b). A reference implementation of SONAR alongside with an example use shown in this paper can be found at Reinecke (2023) and at https://github.com/rreinecke/SONAR.

References

Almeida, S., Holcombe, E. A., Pianosi, F., & Wagener, T. (2017). Dealing with deep uncertainties in landslide modelling for disaster risk reduction under climate change. *Natural Hazards and Earth System Sciences*, *17*(2), 225–241. https://doi.org/10.5194/nhess-17-225-2017

Antoniadis, A., Lambert-Lacroix, S., & Poggi, J.-M. (2021). Random forests for global sensitivity analysis: A selective review. *Reliability Engineering & System Safety*, *206*, 107312. https://doi.org/10.1016/j.ress.2020.107312

Barber, C., Lamontagne, J. R., & Vogel, R. M. (2020). Improved estimators of correlation and R 2 for skewed hydrologic data. *Hydrological Sciences Journal*, *65*(1), 87–101. https://doi.org/10.1080/02626667.2019.1686639

Berghuijs, W. R., Luijendijk, E., Moeck, C., van der Velde, Y., & Allen, S. T. (2022). Global Recharge Data Set Indicates Strengthened Groundwater Connection to Surface Fluxes. *Geophysical Research Letters*, *49*(23). https://doi.org/10.1029/2022GL099010

Best, M. J., Pryor, M., Clark, D. B., Rooney, G. G., Essery, R.. H., & Ménard, C. B., et al. (2011). The Joint UK Land Environment Simulator (JULES), model description – Part 1: Energy and water fluxes. *Geoscientific Model Development*, *4*(3), 677–699. https://doi.org/10.5194/gmd-4-677-2011

Beven, K. J., Almeida, S., Aspinall, W. P., Bates, P. D., Blazkova, S., & Borgomeo, E., et al. (2018). Epistemic uncertainties and natural hazard risk assessment – Part 1: A review of different natural hazard areas. *Natural Hazards and Earth System Sciences*, *18*(10), 2741–2768. https://doi.org/10.5194/nhess-18-2741-2018

Beven, K. J., & Cloke, H. L. (2012). Comment on “Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth’s terrestrial water” by Eric F. Wood et al. *Water Resources Research*, *48*(1). https://doi.org/10.1029/2011WR010982

Bradter, U., Altringham, J. D., Kunin, W. E., Thom, T. J., O’Connell, J., & Benton, T. G. (2022). Variable ranking and selection with random forest for unbalanced data. *Environmental Data Science*, *1*. https://doi.org/10.1017/eds.2022.34

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification And Regression Trees*: Routledge.

Burek, P., Satoh, Y., Kahil, T., Tang, T., Greve, P., & Smilovic, M., et al. (2020). Development of the Community Water Model (CWatM v1.04) – a high-resolution hydrological model for global and regional assessment of integrated water resources management. *Geoscientific Model Development*, *13*(7), 3267–3298. https://doi.org/10.5194/gmd-13-3267-2020

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Conoscenti, C., Ciaccio, M., Caraballo-Arias, N. A., Gómez-Gutiérrez, Á., Rotigliano, E., & Agnesi, V. (2015). Assessment of susceptibility to earth-flow landslide using logistic regression and multivariate adaptive regression splines: A case of the Belice River basin (western Sicily, Italy). *Geomorphology*, *242*, 49–64. https://doi.org/10.1016/j.geomorph.2014.09.020

Denissen, J. M. C., Teuling, A. J., Pitman, A. J., Koirala, S., Migliavacca, M., & Li, W., et al. (2022). Widespread shift from ecosystem energy to water limitation with climate change. *Nature Climate Change*, *12*(7), 677–684. https://doi.org/10.1038/s41558-022-01403-8

Eker, S., Rovenskaya, E., Obersteiner, M., & Langan, S. (2018). Practice and perspectives in the validation of resource management models. *Nature Communications*, *9*(1), 5359. https://doi.org/10.1038/s41467-018-07811-9

ESA. (2010). Global land cover map. Retrieved from http://due.esrin.esa.int/page_globcover.php

Esposito, F., Malerba, D., Semeraro, G., & Kay, J. (1997). A comparative analysis of methods for pruning decision trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(5), 476–493. https://doi.org/10.1109/34.589207

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, *19*(1). https://doi.org/10.1214/aos/1176347963

García-Pedrajas, N. (2011). Evolutionary computation for training set selection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(6), 512–523. https://doi.org/10.1002/widm.44

Gnann, S., Reinecke, R., Stein, L., Wada, Y., Thiery, W., & Müller Schmied, H., et al. (2023a). Functional relationships reveal differences in the water cycle representation of global water models. Preprint (accepted in Nature Water). https://doi.org/10.31223/X50S9R

Gnann S., Reinecke, R. et al. (2023b). Data to "Functional relationships reveal differences in the water cycle representation of global water models" [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7714885

Hanasaki, N., Yoshikawa, S., Pokhrel, Y., & Kanae, S. (2018). A global hydrological simulation to specify the sources of water used by humans. *Hydrology and Earth System Sciences*, *22*(1), 789–817. https://doi.org/10.5194/hess-22-789-2018

Hoch, J. M., Sutanudjaja, E. H., Wanders, N., van Beek, R. L. P. H., & Bierkens, M. F. P. (2023). Hyper-resolution PCR-GLOBWB: opportunities and challenges from refining model spatial resolution to 1 km over the European continent. *Hydrology and Earth System Sciences*, *27*(6), 1383–1401. https://doi.org/10.5194/hess-27-1383-2023

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. https://doi.org/10.1198/106186006X133933

Jiang, S., Bevacqua, E., & Zscheischler, J. (2022). River flooding mechanisms and their changes in Europe revealed by explainable machine learning. *Hydrology and Earth System Sciences*, *26*(24), 6339–6359. https://doi.org/10.5194/hess-26-6339-2022

Jock Blackard. (1998). *Covertype*. https://doi.org/10.24432/C50K5N

Kaur, H., Pannu, H. S., & Malhi, A. K. (2020). A Systematic Review on Imbalanced Data Challenges in Machine Learning. *ACM Computing Surveys*, *52*(4), 1–36. https://doi.org/10.1145/3343440

Koster, R. D., & Milly, P. C. D. (1997). The Interplay between Transpiration and Runoff Formulations in Land Surface Schemes Used with Atmospheric Models. *Journal of Climate*, *10*(7), 1578–1591. https://doi.org/10.1175/1520-0442(1997)010<1578:TIBTAR>2.0.CO;2

Krabbenhoft, C. A., Allen, G. H., Lin, P., Godsey, S. E., Allen, D. C., & Burrows, R. M., et al. (2022). Assessing placement bias of the global river gauge network. *Nature Sustainability*, *5*, 586–592. https://doi.org/10.1038/s41893-022-00873-0

Lee Rodgers, J., & Nicewander, W. A. (1988). Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, *42*(1), 59–66. https://doi.org/10.1080/00031305.1988.10475524

Lesk, C., Coffel, E., Winter, J., Ray, D., Zscheischler, J., Seneviratne, S. I., & Horton, R. (2021). Stronger temperature-moisture couplings exacerbate the impact of climate warming on global crop yields. *Nature Food*, *2*(9), 683–691. https://doi.org/10.1038/s43016-021-00341-6

Loh, W.-Y. (2014). Fifty Years of Classification and Regression Trees. *International Statistical Review*, *82*(3), 329–348. https://doi.org/10.1111/insr.12016

L'vovich, M. I. (1979). *World Water Resources and Their Future*. Washington, D. C.: American Geophysical Union.

MacDonald, A. M., Lark, R. M., Taylor, R. G., Abiye, T., Fallas, H. C., & Favreau, G., et al. (2021). Mapping groundwater recharge in Africa from ground observations and implications for water security. *Environmental Research Letters*, *16*(3), 34012. https://doi.org/10.1088/1748-9326/abd661

Messager, M. L., Lehner, B., Cockburn, C., Lamouroux, N., Pella, H., & Snelder, T., et al. (2021). Global prevalence of non-perennial rivers and streams. *Nature*, *594*(7863), 391–397. https://doi.org/10.1038/s41586-021-03565-5

Müller Schmied, H., Cáceres, D., Eisner, S., Flörke, M., Herbert, C., & Niemann, C., et al. (2021). The global water resources and use model WaterGAP v2.2d: model description and evaluation. *Geoscientific Model Development*, *14*(2), 1037–1079. https://doi.org/10.5194/gmd-14-1037-2021

NASA. (2021). NASA turns to the cloud for help with next generation earth missions. Retrieved from https://www.jpl.nasa.gov/news/nasa-turns-to-the-cloud-for-help-with-next-generation-earth-missions

Pattee, H. H. (1972). Chapter 1 - THE NATURE OF HIERARCHICAL CONTROLS IN LIVING MATTER. In R. Rosen (Ed.), *Foundations of Mathematical Biology* (pp. 1–22). Academic Press. https://doi.org/10.1016/B978-0-12-597201-7.50008-5

Pawlik, M., & Augsten, N. (2015). Efficient Computation of the Tree Edit Distance. *ACM Transactions on Database Systems*, *40*(1), 1–40. https://doi.org/10.1145/2699485

Peterson, T. J., Saft, M., Peel, M. C., & John, A. (2021). Watersheds may not recover from drought. *Science (New York, N.Y.)*, *372*(6543), 745–749. https://doi.org/10.1126/science.abd5085

Pianosi, F., & Wagener, T. (2018). Distribution-based sensitivity analysis from a generic input-output sample. *Environmental Modelling & Software*, *108*, 197–207. https://doi.org/10.1016/j.envsoft.2018.07.019

Puy, A., Roy, P. T., & Saltelli, A. (2022). *Discrepancy measures for sensitivity analysis*. Retrieved from http://arxiv.org/pdf/2206.13470v2

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Reinecke, R. (2023) SONAR (v.0.1). Zenodo. https://doi.org/10.5281/zenodo.10008510

Reinecke, R., Müller Schmied, H., Trautmann, T., Andersen, L. S., Burek, P., & Flörke, M., et al. (2021). Uncertainty of simulated groundwater recharge at different global warming levels: a global-scale multi-model ensemble study. *Hydrology and Earth System Sciences*, *25*(2), 787–810. https://doi.org/10.5194/hess-25-787-2021

Rillig, M. C., Ryo, M., Lehmann, A., Aguilar-Trigueros, C. A., Buchert, S., & Wulf, A., et al. (2019). The role of multiple global change factors in driving soil functions and microbial biodiversity. *Science (New York, N.Y.)*, *366*(6467), 886–890. https://doi.org/10.1126/science.aay2832

Saltelli, A., Jakeman, A., Razavi, S., & Wu, Q. (2021). Sensitivity analysis: A discipline coming of age. *Environmental Modelling & Software*, *146*, 105226. https://doi.org/10.1016/j.envsoft.2021.105226

Sarailidis, G., Wagener, T., & Pianosi, F. (2023). Integrating scientific knowledge into machine learning using interactive decision trees. *Computers & Geosciences*, *170*, 105248. https://doi.org/10.1016/j.cageo.2022.105248

Schaphoff, S., Bloh, W. von, Rammig, A., Thonicke, K., Biemans, H., & Forkel, M., et al. (2018). LPJmL4 – a dynamic global vegetation model with managed land – Part 1: Model description. *Geoscientific Model Development*, *11*(4), 1343–1375. https://doi.org/10.5194/gmd-11-1343-2018

Sheikholeslami, R., & Razavi, S. (2020). A Fresh Look at Variography: Measuring Dependence and Possible Sensitivities Across Geophysical Systems From Any Given Data. *Geophysical Research Letters*, *47*(20). https://doi.org/10.1029/2020GL089829

Shrestha, D. L., Kayastha, N., & Solomatine, D. P. (2009). A novel approach to parameter uncertainty analysis of hydrological models using neural networks. *Hydrology and Earth System Sciences*, *13*(7), 1235–1248. https://doi.org/10.5194/hess-13-1235-2009

Sutanudjaja, E. H., van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H. C., & Drost, N., et al. (2018). PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model. *Geoscientific Model Development*, *11*(6), 2429–2453. https://doi.org/10.5194/gmd-11-2429-2018

Swenson, S. C., & Lawrence, D. M. (2015). A GRACE -based assessment of interannual groundwater dynamics in the C ommunity L and and M odel. *Water Resources Research*, *51*(11), 8817–8833. https://doi.org/10.1002/2015WR017582

Takata, K., Emori, S., & Watanabe, T. (2003). Development of the minimal advanced treatments of surface interaction and runoff. *Global and Planetary Change*, *38*(1-2), 209–222. https://doi.org/10.1016/S0921-8181(03)00030-4

Unwin, A., & Kleinman, K. (2021). The Iris Data Set: In Search of the Source of Virginica. *Significance*, *18*(6), 26–29. https://doi.org/10.1111/1740-9713.01589

Wagener, T., & Kollat, J. (2007). Numerical and visual evaluation of hydrological and environmental models using the Monte Carlo analysis toolbox. *Environmental Modelling & Software*, *22*(7), 1021–1033. https://doi.org/10.1016/j.envsoft.2006.06.017

Wagener, T., Reinecke, R., & Pianosi, F. (2022). On the evaluation of climate change impact models. *WIREs Climate Change*, *13*(3). https://doi.org/10.1002/wcc.772

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., & Schewe, J. (2014). The Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP): project framework. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(9), 3228–3232. https://doi.org/10.1073/pnas.1312330110

West, C., Reinecke, R., Rosolem, R., MacDonald, A. M., Cuthbert, M. O., & Wagener, T. (2023). Ground truthing global-scale model estimates of groundwater recharge across Africa. *The Science of the Total Environment*, *858*(Pt 3), 159765. https://doi.org/10.1016/j.scitotenv.2022.159765

Wright, M. N., Dankowski, T., & Ziegler, A. (2017). Unbiased split variable selection for random survival forests using maximally selected rank statistics. *Statistics in Medicine*, *36*(8), 1272–1284. https://doi.org/10.1002/sim.7212

Zehe, E., & Sivapalan, M. (2009). Threshold behaviour in hydrological systems as (human) geo-ecosystems: manifestations, controls, implications. *Hydrology and Earth System Sciences*, *13*(7), 1273–1297. https://doi.org/10.5194/hess-13-1273-2009