Water Resources Research

RESEARCH ARTICLE
10.1029/2023WR036360

Can eXplainable AI Offer a New Perspective for Groundwater Recharge Estimation?—Global-Scale Modeling Using Neural Network

Hyekyeng Jung[1,2], **Jan Saynisch-Wagner**[3], **and Stephan Schulz**[4]

[1]Department of Ecohydrology, Leibniz Institute of Freshwater Ecology and Inland Fisheries (IGB), Berlin, Germany, [2]Department of Geography, Humboldt University Berlin, Berlin, Germany, [3]Geoforschungszentrum Potsdam, Section 1.3: Earth System Modelling, Potsdam, Germany, [4]Technische Universität Darmstadt, Institute of Applied Geosciences, Darmstadt, Germany

Key Points:

- Estimating groundwater recharge rates at global scale using an ensemble neural network model with 5541 observations and 20 predictors
- XAI can quantify the sensitivity and importance of each predictor, showing non-linearities with long-term precipitation and vegetation index
- Predictions show higher accuracy than the current process-based model, with most behaviors measured by XAI aligning with domain knowledge

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

H. Jung,
hyekyeng.jung@igb-berlin.de

Citation:

Jung, H., Saynisch-Wagner, J., & Schulz, S. (2024). Can eXplainable AI offer a new perspective for groundwater recharge estimation?—Global-scale modeling using neural network. *Water Resources Research*, *60*, e2023WR036360. https://doi.org/10.1029/2023WR036360

Received 3 OCT 2023
Accepted 5 APR 2024

Abstract Due to the difficulties in estimating groundwater recharge and cross-boundary nature of many aquifers, estimating groundwater recharge at large scale has been called upon. Process-based models as well as data-driven models have been established to meet this need. Meanwhile, with the advent of explainable artificial intelligence (XAI) methods, data-driven machine learning models can take advantage of enhanced explainability while keeping the strength of high flexibility. In this study, an ensemble neural network model was built to check the suitability of the model to predict groundwater recharge and the possibility to gain new insights from large data set. Recent large inputs of groundwater recharge data and additional input for the Arabian Peninsula collated in this study were fed to the model with multiple predictors related to climatology considering seasonality, soil and plant characteristics, topography, and hydrogeology. The model showed higher performance (adjusted R^2: 0.702, RMSE: 193.35 mm yr^{-1}) than a recent global process-based model in predicting groundwater recharge. Using XAI methods as individual conditional expectations and Shapley Additive Explanation interaction values, the model behavior was analyzed and possible linear and non-linear relationships between the predictors and the groundwater recharge rate were found. Long-term averaged precipitation and enhanced vegetation index showed non-linear relationships with groundwater recharge rate, while slope, compound topographic index, and water table depth showed low importance to the model results. Most model behaviors followed the domain knowledge, while multi-correlation between predictors and data skewness hindered the model from learning.

Plain Language Summary Estimating groundwater recharge rates at a large scale has been an important task among hydrologists. Both process-based models and data-driven models have been used for this purpose. Despite their high flexibility and high performance, there has been criticism over data-driven models, especially machine-learning models, that the result of the models are difficult to explain. However, new analysis tools called explainable artificial intelligence (XAI) can help explain the model results. In this study, a machine-learning model (ensemble neural network model) has been built at global scale to check if the model can estimate groundwater recharge rates and to check if the model's behavior explained by XAI can give new insights into the processes. Our model shows higher performance compared to a recent global process-based model. XAI tools are used to explain how the model predicted the groundwater recharge rates. Long-term averaged precipitation and enhanced vegetation index show high sensitivity and high importance in predicting groundwater recharge rates, while topographical factors related to slope, curvature, and depth to the groundwater aquifer show low sensitivity and importance.

1. Introduction

Estimating the groundwater recharge rate has been one of the most important tasks for hydrologists since it constitutes one of the key parameters determining the safe yield—the amount of water that is replenished and thus can be sustainably withdrawn (Blöschl et al., 2019; Oki & Kanae, 2006). Currently, rapid decline of groundwater levels has been observed worldwide and accelerated during the past four decades in 30% of aquifers (Jasechko et al., 2024). This is especially alarming in regions where surface water resources are either generally scarce, not constantly available in time and space, or contaminated, so that the population is primarily dependent on groundwater (Schulz et al., 2024).

Many aquifers lie across national boundaries, which require large-scale models for policy decisions. Moreover, groundwater reserves should be considered in the context of global climate change (Gleeson et al., 2021). In this context, hydrological models for analyzing groundwater recharge at the continental or global scale are helpful to consider different interlinkages of climate change and water resource availability. However, regionalization of groundwater recharge on a global scale is challenging, mainly due to the lack of consistent data with high accuracy for groundwater recharge rates. Particularly at larger catchments, estimation of groundwater recharge largely depends on methods under certain assumptions due to the lack of direct measuring methods (Healy & Scanlon, 2010). Moreover, at the large scale, uncertainty still remains regarding the relationships between groundwater recharge and its governing factors (Mohan et al., 2018; Morbidelli et al., 2018; Vereecken et al., 2019). These factors can vary significantly depending on spatial and temporal location and its scales, as the groundwater recharge mechanisms to be considered differ. Also, there are no standardized criteria in practice for assessing the accuracy of estimation (Healy & Scanlon, 2010; MacDonald et al., 2021).

Previously, there have been considerable attempts in developing models to estimate groundwater recharge rates at larger scales. These comprise process-based models (de Graaf et al., 2015; Döll & Fiedler, 2008; Müller Schmied et al., 2021) as well as data-driven models (Berghuijs et al., 2022; MacDonald et al., 2021; Mohan et al., 2018). Usually, process-based models such as land-surface models and hydrological models are targeting on multiple components of the hydrological cycle. However, these models tend to oversimplify the groundwater component, which result in mismatches between field observations and model results (Scanlon et al., 2018; Soltani et al., 2021). Moreover, several model parameters have to be calibrated with a single or a few sets of observation data, which often leads to the problem of non-unique solutions, where the model results in good performance regardless of each component's low fidelity (Beven, 2006; Her & Seong, 2018). On the other hand, there have been data-driven models, which are entirely based on the statistical properties of the data at hand. Different model designs can be applied in this statistical sense. In the study of Mohan et al. (2018), multiple regression models with multiple predictors were used to examine the relationship between predictors and groundwater recharge in a global scale model. This model was based on a data set of 715 sites that were highly skewed to arid and temperate regions. There, multi-collinearity among the predictors was found, but not thoroughly explored in terms of model behavior (Mohan et al., 2018). A linear mixed model and a non-linear regression model with a sigmoid function were applied to build predictive models. Each model performs well at the regional scale (African continent; MacDonald et al., 2021) and global scale (Berghuijs et al., 2022). These studies reassured that groundwater recharge process was firmly constrained by climate factors, especially precipitation. However, the model designs were less flexible since the model structures were predefined before parameters were fitted by the data. Thus, even though other factors such as soil characteristics are expected to affect the groundwater recharge, adding more factors doesn't increase accuracy of the models.

There have been substantial advances of machine learning (ML) in the data science field. For predicting groundwater levels that can be observed relatively easily, ML models have been used successfully at regional scales (Haaf et al., 2023; Wunsch et al., 2022). For the European continent, there is even a groundwater recharge map based on ML, which uses national survey data as training data (Martinsen et al., 2022). However, apart from a few regions with accessible and consistent national survey data, there has been the chronic deficiency of groundwater recharge data. Not only data-driven models but also existing process-based models at the regional and global scale usually have been limited by a relatively small number of field measurements in the subsurface (Seibert et al., 2024) which limits the verifiability and thus constrains the model's credibility. However, recently, comprehensive data sets of ground-based groundwater recharge estimates (i.e., estimates based on field measurements, excluding satellite-based measurements) have been presented by Mohan et al. (2018), Moeck et al. (2020) and MacDonald et al. (2021), which, if used collectively, could address the problem of insufficient data. For large-scale groundwater recharge prediction or its upscaling, various machine learning algorithms can be applied, ranging from classic regression models with low computational cost to computationally expensive neural network (NN) based deep learning models. Thanks to the great increase in computing power in this era, NN models with high computational demands have become more and more common. However, despite their high flexibility NN models have been criticized for being “black boxes” where explaining the model's behavior is difficult. Therefore, great efforts have been made to develop methods for ML model explanation, so-called explainable artificial intelligence (XAI). Using these XAI methods, models with high complexity and high flexibility can also take advantage of high interpretability (Ali et al., 2023).

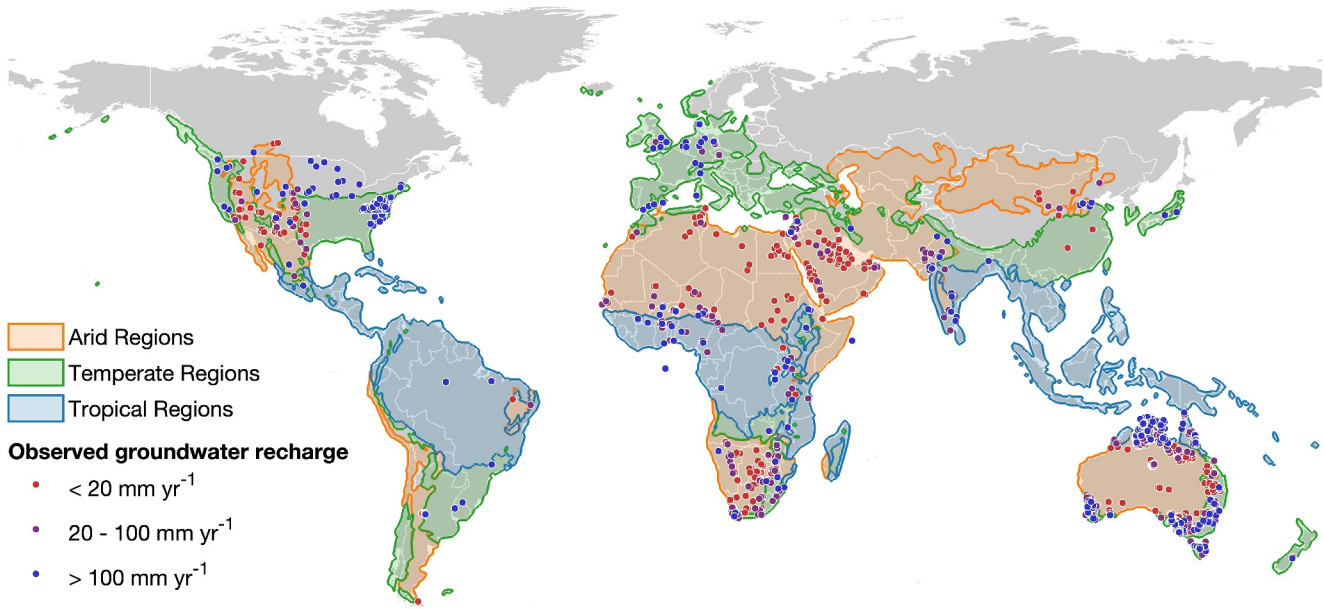

Figure 1. Spatial distribution of groundwater recharge rate estimates (n = 5,541).

Given the recent release of large data sets from the aforementioned meta-studies and the advances of ML models in the data science field, two objectives are addressed in this study: (a) The suitability of the ML model is investigated for the model to predict groundwater recharge rates with unevenly distributed, high dimensional data from ground-based groundwater recharge estimates. This is tested using three different validation strategies (Gleeson et al., 2021): The model results are compared with additionally available observations (observation-based evaluation), with the knowledge of groundwater recharge processes characterized by different climate zones (expert-based evaluation), and with a previous hydrological model (model-based evaluation). (b) To gain new insights from the model, the model behavior is explored at a global scale for different climate zones using XAI. The significance and sensitivity of the predictors and the bivariate interaction effect among the predictors are quantified.

2. Methods

2.1. Data

2.1.1. Collection of Groundwater Recharge Estimates

Estimated groundwater recharge rates have been collated from two global-scale meta-studies (Moeck et al., 2020; Mohan et al., 2018) and from a meta-study for the African continent (MacDonald et al., 2021). Additionally, 92 estimates for the Arabian Peninsula have been collected in this study (Jung et al., 2024). Collected data represent naturally occurring recharge from precipitation. Artificial recharges such as from irrigation return flows and managed aquifer recharge are omitted. If several estimates are available for different periods at one location, the value for the longer period is selected. Studies with a period of less than 1 year are also removed from the data to avoid bias due to seasonal effects in the meta-studies by Mohan et al. (2018) and Moeck et al. (2020). After removing duplicated studies, 5541 samples have been obtained in total. Geographic coordinates as well as groundwater recharge rates from each study have been compiled (Figure 1).

Diverse methods have been used to estimate the groundwater recharge, categorized into water balancing methods, water table fluctuation, and methods using environmental tracers such as the chloride mass balance. The use of different methods to determine recharge rates is known to result in varying estimates (Crosbie et al., 2010). In the data collected by Moeck et al. (2020), tracer methods, which are also the dominant estimation method in other meta-studies, account for about 80%. Nevertheless, the estimation method was not considered in the training data set used in this study. This is due to the lack of relevant information. Many studies did not specify the underlying

Table 1
Description of Predictors

	ID	Predictor	Unit	Resolution	Source
Location	Lat_N	Latitude	degree		
	Long_E	Longitude	degree		
Climatology	KG	Köppen-Geiger climate classification			Beck et al. (2018)
	LTA_P	Long term-averaged Precipitation	mm month^{-1}	0.5°	Harris et al. (2020)
	IAV_P	Intraannual variability of precipitation		0.5°	Harris et al. (2020)
	LTA_T	Long term-averaged Temperature	C°	0.5°	Harris et al. (2020)
	IAV_T	Intraannual variability of temperature		0.5°	Harris et al. (2020)
	LTA_PET	Long term-averaged Potential evapotranspiration	mm day^{-1}	0.5°	Harris et al. (2020)
	IAV_PET	Intraannual variability of Potential evapotranspiration		0.5°	Harris et al. (2020)
Soil and Plant Characteristics	RP	Runoff potential		250 m	Ross et al. (2018)
	EVI	Enhanced Vegetation Index		0.05°	Didan (2021)
	BD	Bulk Density *	g cm^{-3}	5 arc-minute	Global Soil Data Task Group (2000)
	FC	Field Capacity *	mm	5 arc-minute	Global Soil Data Task Group (2000)
	PAWC	Profile available water capacity *	mm	5 arc-minute	Global Soil Data Task Group (2000)
	WP	Wilting point *	mm	5 arc-minute	Global Soil Data Task Group (2000)
Topography	DEM	Elevation	m a.s.l.	3 arc-second	Jarvis et al. (2008)
	SL	Slope	%	3 arc-second	Verdin and Survey (2017)
	CTI	Compound Topographic Index	m	3 arc-second	Verdin and Survey (2017)
Hydrogeology	WTD	Water Table Depth	m	ca. 1000 m	Fan et al. (2017)
	Li	Lithology		0.5°	Hartmann and Moosdorf (2012)

Note. *0–100 cm b.g.l.

method because they obtained data from gray literature in which the estimation method was not specified, or the methods were omitted when collating the data from the previous meta-studies.

2.1.2. Predictors

A large number of factors possibly influencing groundwater recharge were selected and generated. Among them, 20 were finally selected as input predictors for the further neural network development. The selection was based on inspecting the feature importance derived by applying a random forest approach to all possible influencing factors (Breiman, 2001), where all features with non-negligible feature importances were chosen. The final predictors with significant predicting-power for groundwater recharge are listed in Table 1. They can be acquired from publicly available data sets related to climatology, soil characteristics, topography, hydrogeology, and geographic location. Note that the omitted factors are not necessarily unimportant from a hydrological point of view, they are rather redundant from an information-content point of view. Further information on the predictors can also be found in the Supporting Information.

The climate database CRU TS v4.05 includes monthly means of globally interpolated observational data for 1901–2020 (Harris et al., 2020). Both annual mean and monthly mean values are calculated for the 120 years period. Subsequently, the intraannual variability (IAV, Equation 1) of precipitation, temperature, and potential evapotranspiration are calculated.

$$\mathrm{IAV}_f = \frac{\sum_{m=1}^{12} F_m - F}{F} \tag{1}$$

where IAV_f is the intraannual variability of the climate variable f, F_m is the monthly mean of f during the month m, and F is the annual mean of f.

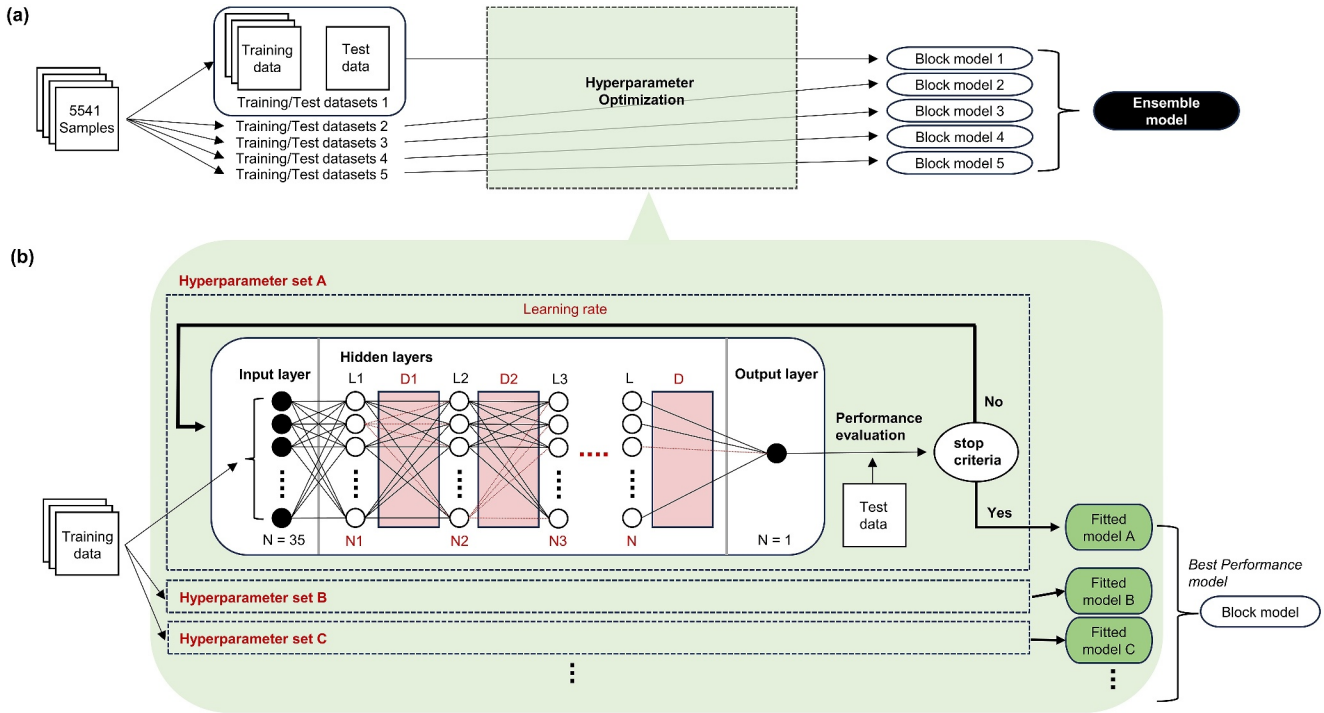

Figure 2. The structure of the (a) ensemble model and (b) the hyperparameter optimization for a set of training and test data.

Also, vegetation as well as soil characteristics affect the percolation process. Therefore, the enhanced vegetation index (EVI) is derived based on MODIS Terra v6.1 satellite surface reflectance imagery (Didan, 2021). Monthly means of EVI between February 2000 to April 2022 are obtained, and data are averaged for the whole period. For the water storage capacity of the soil, relevant features such as bulk density, field capacity, profile available water capacity (PAWC) and wilting point are derived from IGBP-DIS set using a worldwide pedon and soil texture database (Global Soil Data Task Group, 2000). To characterize the topography, elevation as well as its secondary data such as the slope and the Compound Topographic Index (CTI), which is defined as the natural logarithm of a specific catchment area per tangent slope (Verdin & Survey, 2017), are derived from a void-filled digital elevation model (SRTM, Shuttle Radar Topography Mission; Jarvis et al., 2008). Water table depth is defined as the distance between ground level and groundwater table and resulted from inverse groundwater flow modeling (Fan et al., 2017). Runoff potential is extracted from hydrologic soil groups that have been compiled for runoff modeling (Ross et al., 2018).

2.1.3. Preprocessing the Data Set

Predictor values for each groundwater recharge estimate are extracted from gridded or vectorized data based on the coordinate of each sample. Data distributions for each predictor are shown in Figure S1 in Supporting Information S1. Since categorical features cannot be processed by neural networks, they are transformed into numerical vectors of zeros and a single one. The vector length equals the number of classes in a categorical predictor. The position of the one in the vector encodes the specific class of the predictor (one-hot encoding). This results in eight features for "Runoff potential," five features for "Köppen-Geiger climate classification" and 14 features for "Lithology." These 14 lithological classes represent the first level information that is the only one available on a global scale (Hartmann & Moosdorf, 2012). To prevent weighting of a feature based on its scale, all features are standardized by subtracting their mean from the values and by dividing by their standard deviation (Z-scoring).

2.2. Model

Many algorithms exist for regression problems. These include machine learning algorithms such as neural networks (NN), which have been widely applied to numerous geoscientific problems (Reichstein et al., 2019). Their

Table 2
Design and Performance of the Models for Different Training Data Sets With Number of Nodes Per Layer (L) and the Frequency of Dropout (D)

	Model design						Predicted values	Performance	
Block model	Input	L1	D1	L2	Output	Learning rate	Mean ± standard deviation [mm yr^{-1}]	Adjusted R^2	RMSE [mm yr^{-1}]
1	35	10	0		1	0.01	203.08 ± 300.03	0.685	193.91
2	35	40	0.2		1	0.01	234.12 ± 328.13	0.672	197.93
3	35	70	0		1	0.001	200.00 ± 279.46	0.680	197.92
4	35	70	0.2	70	1	0.01	180.63 ± 253.53	0.660	201.68
5	35	10	0		1	0.01	187.39 ± 267.37	0.671	198.18

advantages are that they are universal, adaptive, and can well handle high-dimensional nonlinear problems (Irrgang et al., 2021). Therefore, in our approach, a feed forward neural network is chosen for estimating groundwater recharge rates.

Five different NN models (hereafter referred to as "block models") are built from five sets of randomly split training and test data with a ratio of 8:2 (cf. Chollet, 2017; Krizhevsky et al., 2017). Subsequently, the results of the five block models are averaged and hereafter referred to as ensemble model (Figure 2a). This strategy increases the ensemble model's robustness and prevents overfitting (e.g., Breiman, 2001). The hyperparameter-configuration (i.e., number of hidden layers, number of neurons each layer, learning rate and dropout) of each block model is optimized by a hyperparameter search within the Tensorflow/Keras environment (Figure 2b, Abadi et al., 2016). To increase the model's stability and to favor generality during the model learning, weak signals are neglected by not being conveyed to nodes in the next layer through ReLU (Rectified Linear Unit) activation functions (Chollet, 2017). An option for dropout is used in all layers. The results can be seen in Table 2. In general, all input layers consist of the 20 predictors (Table 1) followed by a first layer of 35 nodes. An output layer of one node was set for the groundwater recharge rate estimate.

2.3. Model Explanation

Two complementary XAI methods, namely aggregated individual conditional expectation (ICE, Goldstein et al., 2015) and Shapley additive explanations (SHAP, Lundberg & Lee, 2017), are used to tackle the "black box" nature of the neural network operator. The two XAI methods are based on different approaches and theories and are mutually supportive to each other's limitations. Aggregated ICEs are used to find out linear and non-linear relationships between the predictors and the groundwater recharge rate estimates and to measure the respective sensitivity within the neural network. SHAP measures the importance of a single predictor based on the performance decline of the model without this predictor. To measure this decline in a computationally feasible way, a predictor is randomly replaced by typical values of itself. Subsequently, the average performances of the models with the original values and the random values are compared. This is repeated for each predictor.

While aggregated ICEs are good at describing the non-linear relationships and the sensitivity of the groundwater recharge rates with respect to the predictors, the method has limitations since independent predictors are assumed but not guaranteed. This limitation is lifted in SHAP analysis as this method explicitly includes interactions between the predictors. At the same time, the limitation of SHAP is that it does not, in contrast to ICEs, approximate the model's behavior outside the range of the data set.

ICEs and SHAP values have been grouped according to the climate zone, and model behavior depending on the climate zone has been analyzed. Due to the small number of samples for cold regions and polar regions, only results in tropical regions, arid regions, and temperate regions are meaningful to explain and discussed mainly.

3. Results and Discussions

3.1. Model Performance and Stability

The ensemble NN model has provided predictions of the groundwater recharge rates across the globe at a resolution of 0.25° × 0.25° (Figure 3a). The distribution of global groundwater recharge rates generally aligns with Köppen-Geiger climate classification and the trend that has been reported in the previous data-based model

10.1029/2023WR036360

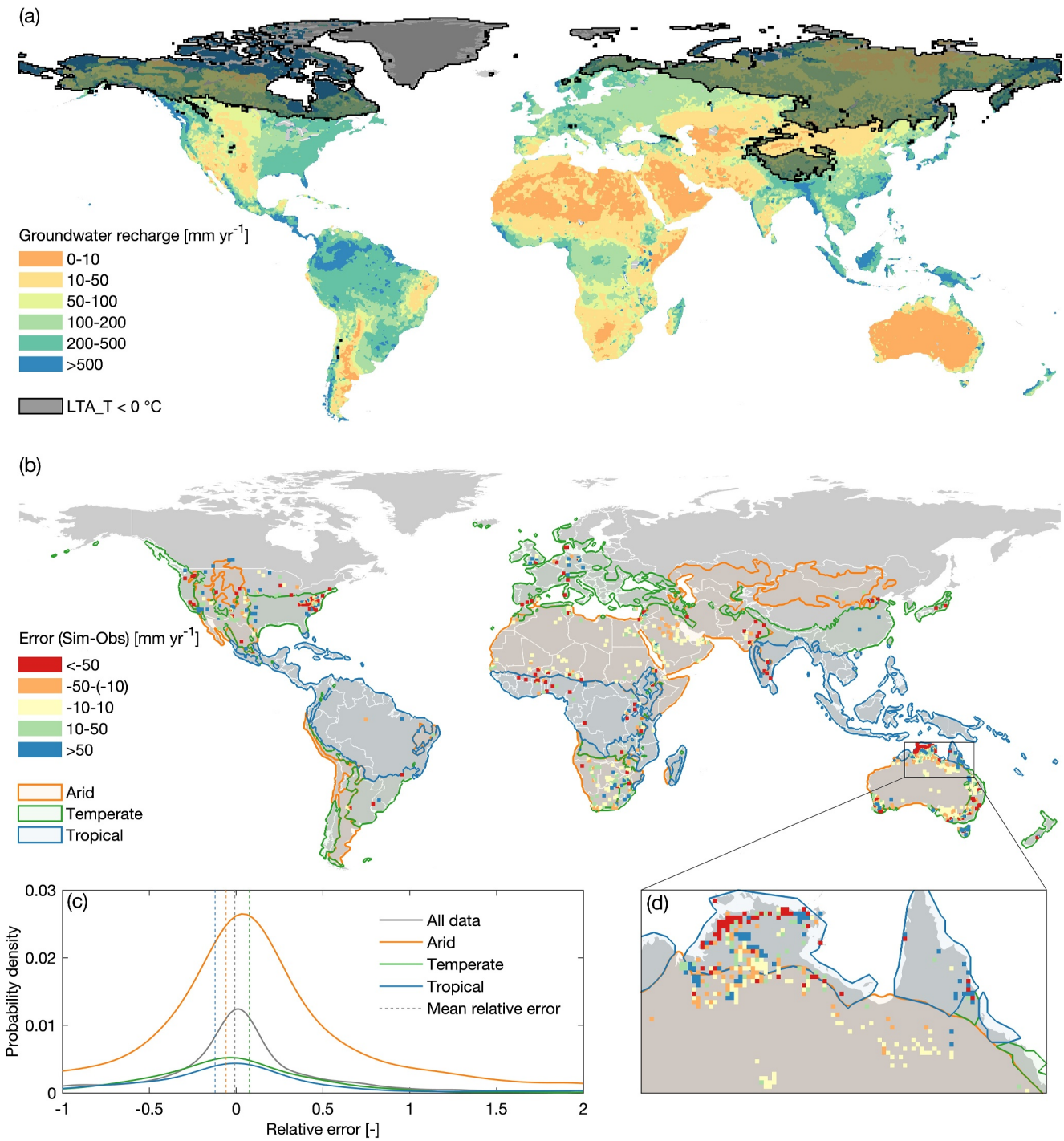

Figure 3. Groundwater recharge predictions (a); model error (residuals) as average values on a 0.25° grid for the entire world, filtered with a 3 × 3 median filter for better visibility (b), and northern Australia (d); probability density functions of the relative errors for the different climate zones (c).

(Berghuijs et al., 2022) and process-based model (Müller Schmied et al., 2021). However, notable discrepancies are identified, particularly in the polar regions, where the ensemble NN model predicts significantly higher recharge rates. This deviation can be attributed to the absence of observation data in permafrost regions, especially in northern latitudes. The model behavior that have led to high groundwater recharge rates in permafrost areas are able to be explained with the support of XAI. The small number of observations in polar regions (3

Table 3
Performance of the Ensemble Model

					Performance		Stability	
		Number of samples	True mean	Predicted mean	Adjusted R^2	RMSE	Error reduction	Standard deviation of block models
Climate zone	Tropical	2403	439.69	404.08	0.59	279.26	0.34	53.01
	Arid	2022	24.20	13.41	0.13	60.62	1.13	5.79
	Temperate	983	116.74	92.79	0.55	98.79	5.82	15.97
	Cold	73	166.57	176.55	0.52	89.67	4.31	36.77
	Polar	3	119.97	111.48	−3.87	9.53	57.41	57.90
Lithology	Unconsolidated sediment (su)	3488	285.94	259.22	0.69	226.13	0.56	33.15
	Mixed sedimentary rock (sm)	579	53.79	35.43	0.22	91.61	3.16	10.27
	Metamorphic (mt)	442	251.31	216.46	0.61	174.49	1.26	35.40
	Siliciclastic sedimentary rock (ss)	420	108.54	96.25	0.67	91.83	3.26	23.23
	Carbonate sedimentary rock (sc)	239	113.14	100.20	0.63	110.45	2.02	21.83
	Basic volcanic rock (vb)	108	45.89	37.23	0.71	44.67	5.91	11.44
	Acid volcanic rock (va)	76	32.52	29.46	0.50	35.53	−7.11	8.75
	Acid plutonic rock (pa)	63	96.52	94.88	0.50	98.98	1.93	24.79
	Intermediate volcanic rock (vi)	21	141.59	85.60	0.28	130.23	9.23	16.26
	Basic plutonic rock (pb)	17	67.52	62.69	0.93	21.01	−3.28	19.17
	Pyroclastic (py)	13	43.72	11.74	0.19	131.50	12.48	17.46
	Evaporite (ev)	7	6.22	10.35	−4.31	20.61	−4.75	34.99
	Water bodies (wb)	6	123.79	115.84	0.92	50.23	1.34	18.93
	Intermediate plutonic rock (pi)	5	314.97	365.07	0.34	264.28	13.44	149.07
Continent	Oceania	4564	247.24	222.51	0.70	205.85	0.61	30.80
	North America	331	180.29	159.53	0.58	114.92	7.27	22.41
	Africa	303	56.23	41.19	0.28	102.07	−0.38	13.96
	Asia	198	76.36	56.51	0.31	116.49	4.14	17.38
	Europe	54	62.54	61.53	0.63	63.02	95.22	26.34
	South America	34	192.51	156.77	0.05	152.91	−90.24	16.16

Note. All units are "mm yr^{-1}" except number of samples and adjusted R^2.

samples) leads the predictor representing polar climate to have a high weight to compensate weak representation during the learning process. On the other hand, all three observations classified to polar area are located in a mountain near the Alpes, where the groundwater recharge rate is around 100 mm/yr. Thus, the model has trained itself radically with other several predictors to offset the high weight of the predictor for polar regions, which can be seen in the aggregated ICEs of polar region (Figure S9 in Supporting Information S1). Unlike the model behavior in the other climate regions, the model has related low EVI with high groundwater recharge rate and shown high sensitivities in elevation and slope. This indicates the risk that the AI model works as a "Mathematical Marionette" with scarce observation. In view of the limited availability of corresponding observational data, areas where the long-term averaged temperature is below 0°C are marked in Figure 3a.

The predicted values of the ensemble model have been compared with the observed values (Figure 3b). On average of all samples, the adjusted R^2 score is 0.702 and the Root mean squared error (RMSE) is 193.35 mm yr^{-1} with ≤25 mm yr^{-1} in 51% of samples and ≤100 mm yr^{-1} in 75% of samples. The model's error tends to be high in regions where high groundwater recharge has been observed, indicating a relative error of our model results rather than an error with a constant magnitude (Figure 3c).

The performance of the ensemble model has been scrutinized in different subgroups of data according to the climate zone, location and lithology (Table 3). Since the data is highly skewed to few dominant properties in most predictors (Figure S1 in Supporting Information S1), it might be expected that there is a higher chance for the

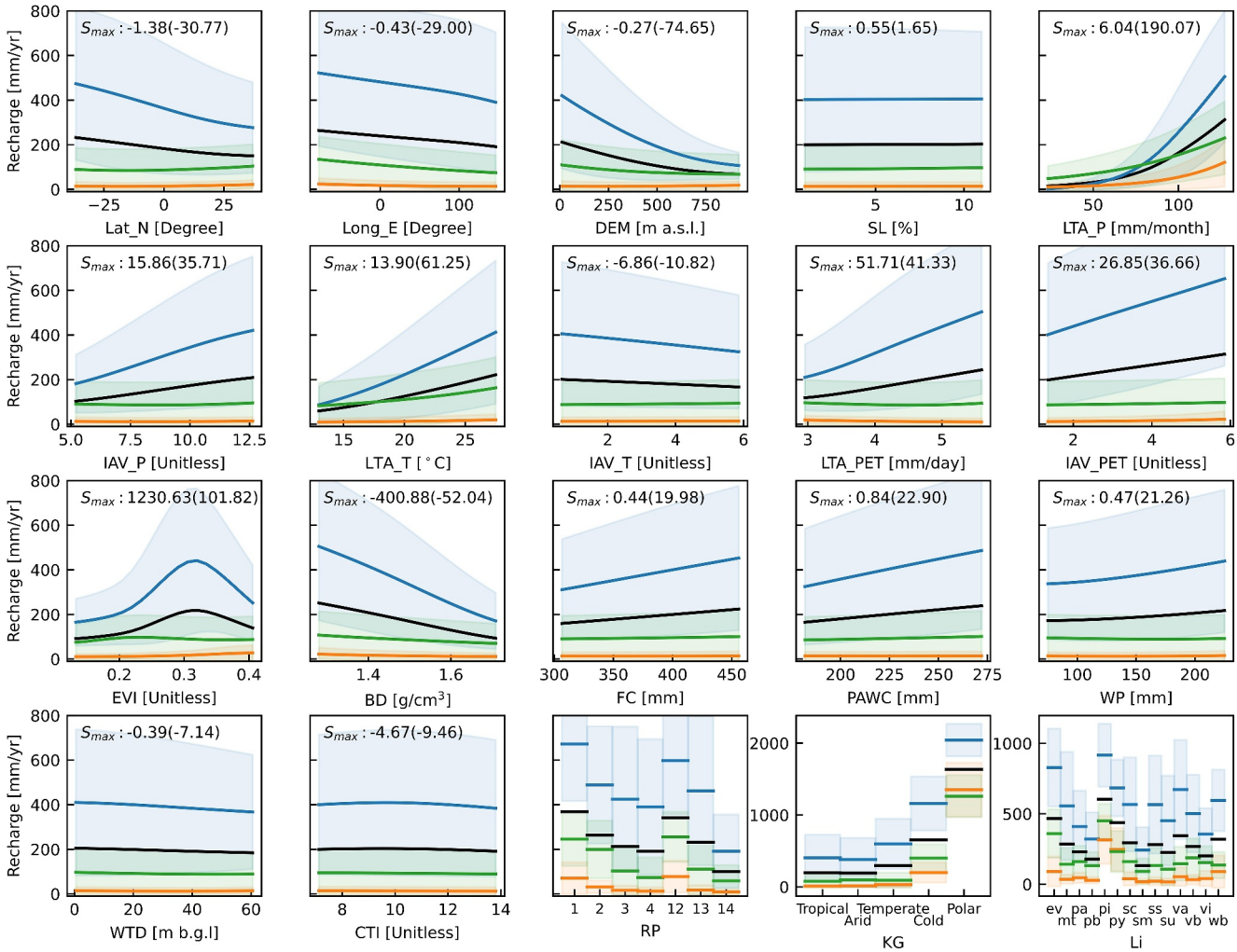

Figure 4. Individual conditional expectations of each predictor are aggregated by climate zone (Blue: Tropical, Yellow: Arid, Green: Temperate, Black: Average of all climate zones). 5 to 95 percentile of predictor is displayed. 95% confidence interval is shown around the average line over climate zones. Sensitivities are represented in each plot by maximum slope S_{max} (mm yr^{-1} per unit of predictor) and normalized S_{max} in brackets (no unit).

model to be fit to only dominant properties. However, the model has shown relatively weak performances particularly over arid regions despite its large number of samples (cf., Oceania in Figure 3d and Table 3). The reason for this can be that, in contrast to the tradeoff often observed in classical regression approaches, a NN can fit several different relationships without jeopardizing each other. Consequently, one of the reasons of high RMSE despite a large data fraction can be that the observational errors, that is, measurement error and data uncertainty, in ground-based estimates of groundwater recharge as well as uncertainties of the predictors are relatively high compared to the scale of observed groundwater recharge in these arid regions.

The block models are established from 5 randomly split sets of training and test data. No big difference among the averaged performances of the block models derived from different train and test data sets has been found, which implies general stability of the model result (Table 2). Compared to the block model 1 (the block model of the best performance), the model's error decreases on most of the properties in the ensemble model, and improvement on less representative properties tends to be larger than the dominant properties.

3.2. Model Sensitivity and Non-Linear Behavior

The linearity of the relationship between each predictor and predicted groundwater recharge rate has been analyzed as well as the sensitivity of each predictor (Figure 4). It results that the two most sensitive predictors (normalized

slope >100), long-term-averaged precipitation (LTA_P) and enhanced vegetation index (EVI), show non-linear relationships with groundwater recharge rate. Groundwater recharge rate increases exponentially as the average precipitation increases. This follows the domain knowledge that groundwater recharge rate can be estimated by a power function of the precipitation (Keese et al., 2005). The non-linear behavior in tropical regions between vegetation and groundwater recharge rate is shown by a positive relationship up to an EVI of about 0.3 and a subsequent drop in groundwater recharge as EVI increases. Interestingly, the sensitivity of EVI is greatly weaker in arid and temperate regions than in tropical regions. A possible explanation might be that distinctive vegetation is also an indication of sufficient rainfall. However, in the case of very dense vegetation (such as in tropical rainforests), at one point large interception losses caused by the dense canopies as well as uptake from plants reduce the water amount available for deep percolation and thus causing a negative effect on groundwater recharge. Such a relationship, characterized by an initial increase and then from intermediate tree cover on a subsequent decrease in groundwater recharge with an increase of canopy cover in the tropics, is also described by Ilstedt et al. (2016).

The other long-term averaged climate predictors are temperature (LTA_T) and potential evapotranspiration (LTA_PET). Both show high sensitivities and a fairly linear relationship with the groundwater recharge. Especially in the tropics, the groundwater recharge rate increases as temperature and potential evapotranspiration increase, which fits the domain as high temperatures, facilitating also high potential evapotranspiration, are associated with particularly high rainfall amounts in these environments (Adler et al., 2008). Also, the seasonality of weather variables, represented by the intraannual variability of climate predictors, can be a factor in this context as well. While the sensitivity of intraannual variance of temperature (IAV_T) is very low, IAVs of precipitation (IAV_P) and potential evapotranspiration (IAV_PET) show moderate sensitivities on the model. Especially, the sensitivity of the model to IAV_P is relatively high in tropical regions, despite that groundwater recharge has been known to have a seasonal trend outside the tropical region (Healy & Scanlon, 2010). Yet, it might support a recent finding that groundwater recharge in tropical regions is also more likely to occur from single heavy rainfall events rather than light rain (Jasechko et al., 2014; Jasechko & Taylor, 2015).

Another interesting observation is that the model is insensitive to predictors that have been expected to be relevant factors for groundwater recharge. This is noticed for some predictors related to topography and hydrogeology, which are slope (SL), compound topographic index (CTI), and water table depth (WTD). One possible explanation for the low sensitivity of the WTD variable could be that this variable can act in different directions. On the one hand, it is to be expected that high groundwater recharge will also tend to lead to high groundwater levels. On the other hand, high groundwater levels can also have a negative impact on the net groundwater recharge. This is the case when the groundwater water table is above the zero-flux plane, allowing capillary rise and root water uptake from groundwater. In addition, other predictor variables such as elevation or lithology, which can also be assumed to have an influence on WTD, could mask the effect of WTD on groundwater recharge. Accordingly, high sensitivities with linear relationships to groundwater recharge are shown by elevation (DEM). Predictions of groundwater recharge increase as elevation decreases. This behavior has been also observed in a study on groundwater recharge estimation for Europe using machine learning by Martinsen et al. (2022). The direction of the model behavior with respect to elevation, as well as slope (despite the lower sensitivity), suggests that precipitation tends to infiltrate in the plain lowlands, as expected (Jaafarzadeh et al., 2021). Like elevation, bulk density (BD) also shows a relatively high sensitivity and is negatively correlated with groundwater recharge, which is similarly observed by Martinsen et al. (2022). At first, this seems counterintuitive from a physical point of view, since finer materials show usually lower bulk densities than coarser ones and that those coarser materials are characterized by higher hydraulic conductivities, which favor percolation. Contrary to this assumption, however, there are also observations that it can be the other way around. For example, Price et al. (2010) showed in a study on soil hydraulic properties that bulk density can also be negatively correlated with hydraulic conductivity. The explanation probably lies in the fact that the processes and relationships in natural environments are complex and that a direct conclusion about one specific soil characteristic is not readily possible. In fact, bulk density can be partially correlated with porosity and thus water holding capacity, which also have an influence on groundwater recharge. This might also explain why other soil characteristics related to water holding capacity show relatively low sensitivity on the model, which are field capacity (FC), wilting point (WP), and PAWC. Surprisingly, they all show a positive effect on groundwater recharge, although their relationship to each other (PAWC = FC—WP) actually does not allow this. A possible explanation could be given by their low sensitivity caused by being confounded with bulk density. Interaction effect of bulk density and water holding capacity is described in a subsequent Section 3.3.2.

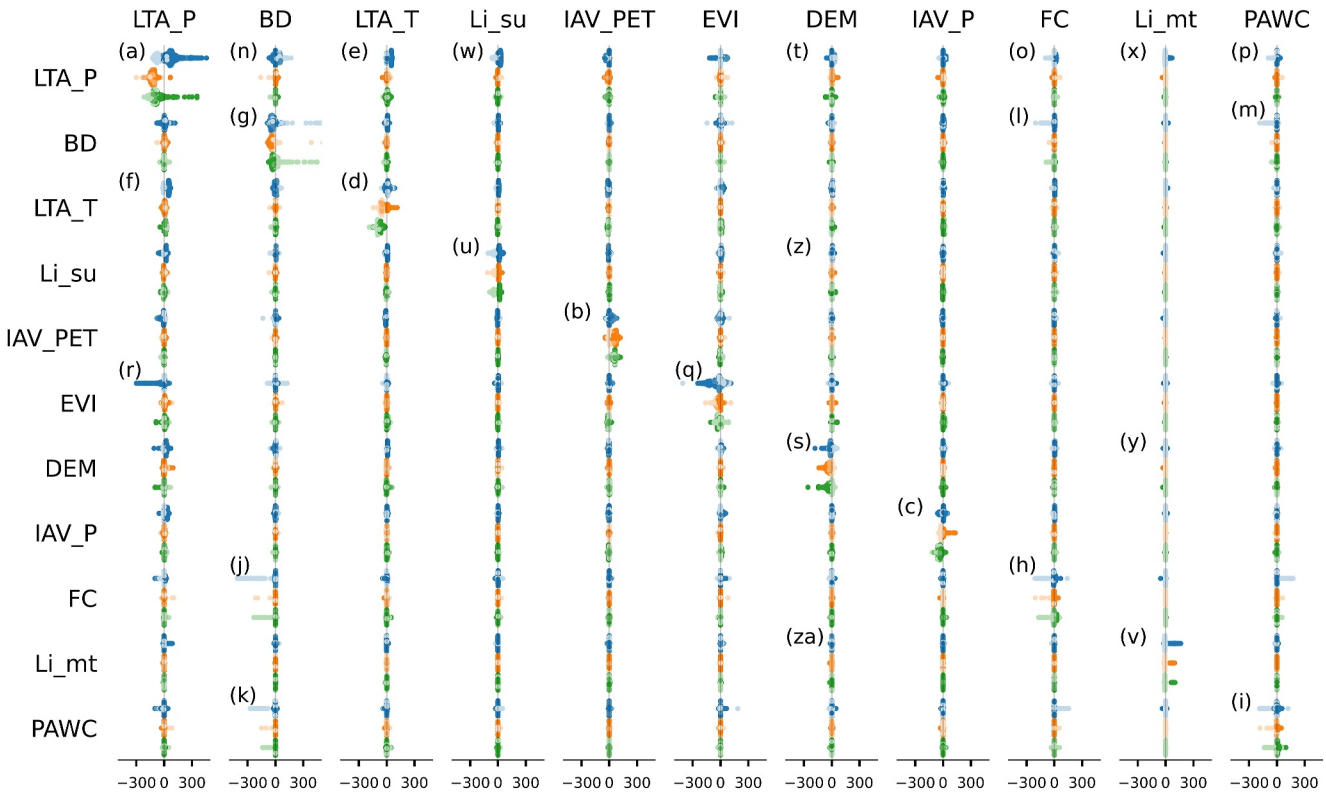

Figure 5. Highlight of main and interaction effects in tropical (blue), arid (yellow) and temperate (green) region. Selected features are displayed in the order of absolute value of SHAP. Feature value of the row is illustrated by lightness of color (dark: high value, light: low value). SHAP interaction values for more predictors can be seen in Figures S10, S11, and S12 in Supporting Information S1.

For category-type predictors, such as Köppen-Geiger climate classification (KG), Lithology (Li), and Runoff potential (RP), predicted recharge rates depend on the proportion of samples within a particular class rather than following the domain knowledge. For example, the model sensitivities are high with rare properties such as the climate zones cold and polar climate or the lithologies evaporate and intermediate plutonic rock (Table 3). This can be due to the one-hot encoding and subsequent normalization that often result in extremely high input values for rare properties. This weighting of rare properties could hinder the model from learning the process related to the categorical predictor if the data distribution is highly skewed. Thus, it should be carefully decided whether one-hot encoding and normalization are to be applied on skewed categorical data, depending on the model purpose.

3.3. Quantifying Contribution of Predictors

SHAP values, which represent a measure of the contribution of a predictor on the model's prediction, have been calculated for each predictor (Figure S2 in Supporting Information S1) and decomposed to a main effect (diagonal) and interaction effects (off-diagonal) in Figure 5. The tendency of SHAP values depending on a predictor in the samples describes the relationship between groundwater recharge and the predictor solely (main effect) or interaction with the other predictors (interaction effect). The magnitude of SHAP values is a measure of the importance of each predictor under the presumption that the predictor of high SHAP (large contribution to the model) has a high influence on prediction (Figures S4, S5, and S6 in Supporting Information S1).

3.3.1. Climatology

A clear positive relation between long-term averaged precipitation (LTA_P) and groundwater recharge can be observed for tropical and temperate regions (Figure 5 (a)). In arid regions, however, long-term averaged precipitation is not a determinant predictor and not even shows a clear tendency in the direction of impact in the ensemble model. In the neural network model that is only fed by the samples from arid regions (Figure S7 in Supporting Information S1), however, a clearly positive but weaker relationship of long-term averaged

precipitation and groundwater recharge is observed as the fifth important predictor following vegetation, latitude, intraannual variance of precipitation, and potential evapotranspiration.

While seasonality (represented by the IAVs) of potential evapotranspiration (Figure 5 (b)) and precipitation (Figure 5 (c)) positively affect groundwater recharge predictions of the ensemble model in arid and temperate regions, the seasonal effect is relatively weak in both main and interaction effects in tropical regions. This corresponds to the domain knowledge, as groundwater recharge tends to follow seasonal weather fluctuations, which however are not or less pronounced in the tropics (Healy & Scanlon, 2010; Mohan et al., 2018).

Regarding long-term averaged temperature (LTA_T), its main effect is not significant in tropical regions, but as temperature increases, the main effect increases in the direction of increasing groundwater recharge in arid and temperate regions (Figure 5 (d)). In tropical regions, the interaction effect of precipitation and temperature is stronger than the main effect of temperature (Figure S4 in Supporting Information S1), and an increase in these predictors also leads to an increase in the predicted groundwater recharge when both precipitation and temperature have high values (Figure 5 (e), (f)). This might indicate that the model uses the temperature predictor at high precipitation rates as an indicator for tropical regions rather than following the expectation that groundwater recharge decreases with higher temperature, as this usually also results in higher evapotranspiration.

3.3.2. Soil Characteristics and Vegetation

First, regarding the main effects driven by one predictor, as bulk density (BD) is low, and as field capacity (FC) and PAWC is low, the main effects of each decrease the prediction in all climates (Figure 5 (g), (h), (i)). Also, for all climates, low bulk density decreases the interaction effects of water holding capacity (Figure 5 (j) and (k)), and vice versa (Figure 5 (l) and (m)). However, related to the interactions of bulk density and water holding capacity (FC and PAWC) with precipitation, they show different behaviors by climate. In tropical regions, in case of low bulk density and high water holding capacity, the interaction effect of precipitation changes the prediction positively, while it affects it negatively in temperate and arid climates (Figure 5 (n), (o), (p)). From this, it can be concluded that the effect of precipitation is affected by other predictors of soil characteristics, that is, precipitation effect decreases the predicted groundwater recharge rates if the soil has low bulk density and high water holding capacity in tropical regions and if high bulk density and low water holding capacity in temperate and arid regions.

Vegetation has been considered an important factor for groundwater recharge (Kim & Jackson, 2012; Moeck et al., 2020; Mohan et al., 2018). In this study, the predictor for vegetation (EVI) is more important in tropical and arid regions, but less important in temperate regions (Figure S2 in Supporting Information S1). However, the main effect of vegetation doesn't show a clear direction in tropical and temperate regions, but positive effect of vegetation is observed in arid region (Figure 5 (q)). In tropical regions, the effect of vegetation by precipitation decreases the prediction as long-term averaged precipitation is high (Figure 5 (r)).

3.3.3. Topography, Hydrogeology, and Lithology

At a global scale, orographic effects can be related to elevation. The main effects in all three climates increase as the elevation (DEM) becomes low (Figure 5 (s)), which may imply that the groundwater recharge increases in lower elevations. However, the interaction effect shows differences depending on the climate zone. In tropical regions, a positive interaction effect with precipitation is found as elevation is lower. In contrast, in arid regions, the interaction effect of precipitation decreases as elevation is getting lower (Figure 5 (t)).

The main effect of DEM might be explained by the mountain front recharge scheme consisting two mechanisms: mountain-block recharge, which occurs through the faults and fractures of the mountain bedrock, and surface mountain front recharge that occurs at the foot of the mountain by accumulated mountain-originated surface water (Markovich et al., 2019). Mountain front recharge is known to be dominant in arid regions (Markovich et al., 2019; Scanlon et al., 2006). Due to the orographic effect, the mountainous areas in arid environments receive a relatively large amount of precipitation (Whitford & Duval, 2020). Although precipitation often cannot infiltrate directly onto the slopes due to high rainfall intensity and the water-repellent effect of dry soils in arid regions, surface runoff collects in depressions and valleys (wadis), where it leads to increased groundwater recharge. Moreover, interaction effects of precipitation and elevation might imply that mountain block recharge can also engage significantly in arid regions. These interpretations are also underpinned by the topographical distribution of the training data (Figure S3 in Supporting Information S1). For example, within arid environments,

the sites in the mountain regions show about twice as much annual groundwater recharge (43 mm, $n = 137$) as in the lowlands (23, $n = 1893$). For the training data in the other climate zones, the situation is exactly the opposite. In the mountain regions, the average annual groundwater recharge (140 mm, $n = 180$) is only about half as high as in the lowlands (359, $n = 3166$).

The predictors related to surface morphology give mere contributions to groundwater recharge prediction. Especially, slope and curvature have been reported to be sensitive predictors in domain knowledge and local and regional-scaled studies (Jaafarzadeh et al., 2021; Morbidelli et al., 2018). However, at large scales, the role of the slope has been still quite unclear and controversially discussed (Morbidelli et al., 2018; Moeck et al., 2020; Mohan et al., 2018; Vereecken et al., 2019). This might imply that slope and curvature constitute an important predictor for groundwater recharge at a small scale, but not on a larger scale, where elevation itself is of greater relevance. In addition to slope, the depth to water table affects weakly the model of this study, although it has been reported to have an impact on groundwater recharge (Carrera-Hernández et al., 2011; Moeck et al., 2020).

Unconsolidated sediments (Li_su) and metamorphic rocks (Li_mt) show meaningful importance among the boolean lithology predictors. The main effect increases groundwater recharge rates for unconsolidated sediments or metamorphic rocks in all climates (Figure 5 (u), (v)). However, the interaction effect shows differences between climates. In tropical regions, the interaction effect of long-term averaged precipitation tends to incline in case of prevailing unconsolidated sediments and metamorphic rocks. Contrarily, in arid and temperate regions, it tends to decline for unconsolidated sediments or metamorphic rocks (Figure 5 (w), (x)). Also, in arid and temperate regions, the interaction effect of elevation decreases the prediction for metamorphic rocks (Figure 5 (y)). On the other hand, contributions of lithology are affected by elevation. As elevation is low, the interaction effect of unconsolidated sediments is negative (Figure 5 (z)) and the interaction effect of metamorphic rocks is positive (Figure 5 (za)). While it can be assumed that unconsolidated sediments have a favorable effect on infiltration, it's expected that this would be not the case for metamorphic rocks. However, the model reacts opposite with the main effect for metamorphic rocks as expected. The prediction of the groundwater recharge rate depends more on interactions with other predictors rather than on a single predictor.

3.4. Causality and Comparison to Process Based Models

XAI can discover the associations required for the predicting model, but does not show necessarily causality. In this study, also associations of predictors that do not correspond to the expected behavior have been found. For example, the direction of the contributions of climate predictors, such as long-term averaged temperature and potential evapotranspiration, and the directions of the contributions of field capacity, profile available water, and wilting point are the opposite of the domain knowledge for some climate zones. This could be due to the fact that the related predictors were inter-correlated and confounded. Especially high multi-correlation has been found between the climate predictors and between field capacity and wilting point (Figure S8 in Supporting Information S1).

In the case of the aforementioned climate predictors, a simple conceptual causal model can be introduced to explain this. In the conceptual causal model, temperature increases potential evapotranspiration, both of which have negative impacts on groundwater recharge since potential evapotranspiration decreases the amount of available water. However, in the model of this study, precipitation, the strongest predictor, is not independent on both. For example, high precipitation is spatially associated with high temperature and high evapotranspiration in the tropics. Since the positive effect of precipitation is mathematically much stronger compared to the negative effect of temperature and potential evapotranspiration on groundwater recharge, groundwater recharge rate can seem to be increased as temperature and potential evapotranspiration increase. Thus, these two predictors could be considered in the model as positive factors in groundwater recharge. Especially, in this exemplary causal model, even though the temperature doesn't have a direct relation to groundwater recharge, it can be considered an effective predictor affecting groundwater recharge positively. Not only this “confoundedness” issue but redundancy among the predictors (e.g., field capacity and wilting point or intraannual variance of temperature and evapotranspiration) might cause underestimation of feature importance. These issues have been noticed and also expected in this study, due to the multi-correlation among the predictors observed in reality. Thus, confoundedness and redundancy are difficult to avoid in data-driven groundwater recharge estimation. Also, the set of predictors cannot be guaranteed to be complete. Not all relevant features for estimating groundwater recharge at a global scale are clearly determined. Even if the complete set of predictors would be determined, not all

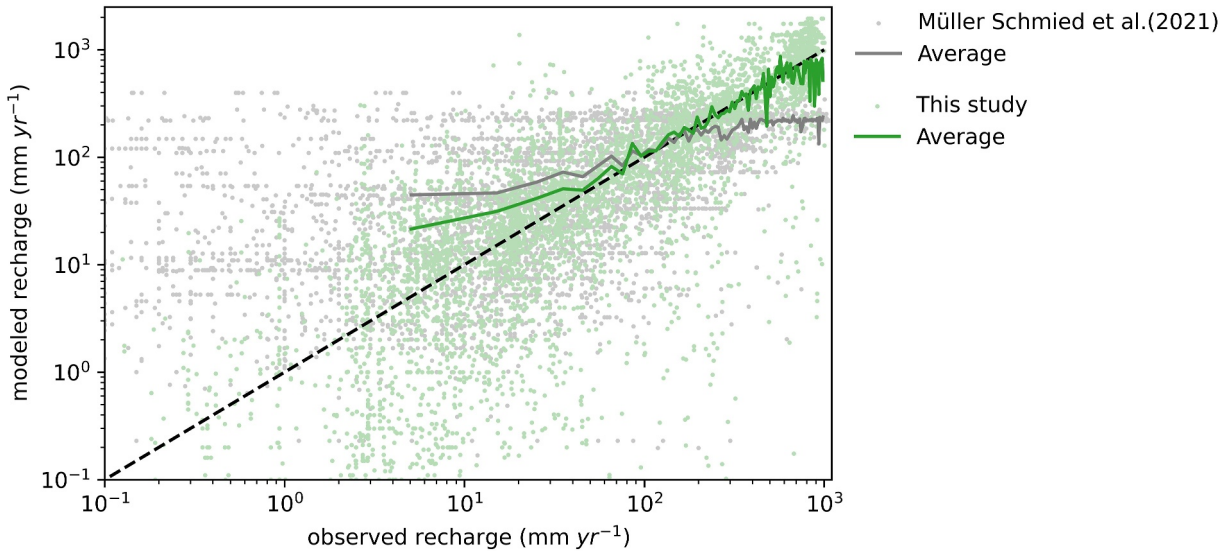

Figure 6. Pointwise comparison of modeled groundwater recharge rates from this study and Müller Schmied et al. (2021) with observations, based on the collected data of meta studies and own literature survey (cf. 2.1.1.). Aggregating is carried out over an interval of 10 mm y^{-1}.

features are measurable. For example, groundwater withdrawal can be an important factor (Shamsudduha et al., 2011), but due to the lack of proper monitoring networks (or their availability) in large parts of the world, it is difficult to obtain reliable spatially and temporally continuous data on it.

Moreover, there are limitations that ground-based data inherently have. For example, uncertainties caused by different spatiotemporal assumptions of estimating methods, Improper application of methods due to lack of site-specific knowledge or expense, non-existence of standard to evaluate the accuracy of different methods, and so on (Healy & Scanlon, 2010). Also, the size and skewness of the data set for groundwater recharge can be a limitation in the model learning process. Groundwater recharge is known to be a complicated process related to many variables. With increasing dimensionality, where the amount of data required to cover the variability of predictors increases exponentially, capturing all the relationships requires a considerably large amount of data as the number of variables increases. Besides, skewness can largely hamper covering the variability of predictors. Additionally, observational errors in ground-based estimating methods may make the model difficult to learn. Especially, assumptions regarding the temporal and spatial scale of the estimation are different according to the method, which can lead to discrepancies between the estimates. For example, in regions where infiltrated water can take up to geological time scales to cause a groundwater system response (Cuthbert et al., 2019), the recharge rate can be greatly misestimated depending on assumptions and spatiotemporal scales of methods (Moeck et al., 2020; Yenehun et al., 2022).

Despite these limitations, the performance of the data-driven model is comparable to conventional process-based hydrological models for predicting groundwater recharge rates on a global scale, even though the model has not been built specifically for high performance. Compared to the estimated naturally occurring groundwater recharge from precipitation of the recently developed process-based hydrological model WaterGAP Global Hydrology Model v2.2d (Müller Schmied et al., 2021), the model of this study shows even better scores in common performance criteria for the prediction of the groundwater recharge rate (Figure 6). Root mean squared error (RMSE) and adjusted R^2 score are 328 mm yr^{-1} and 0.14 for WaterGAP, and 193 mm yr^{-1} and 0.70 for the ensemble NN model in this study, respectively. The data-driven model of this study shows its strength in lower variance and in a better fit in predicting the groundwater recharge at higher recharge rates. High performance for higher rates was also observed in another data-driven model (Berghuijs et al., 2022), where moving average over 10% of the data is used for model comparison (RMSE: 218 mm yr^{-1} in data-driven model while WaterGAP model shows 141 mm yr^{-1}). In the study of Berghuijs et al. (2022), three other process-based models are additionally compared and show similar results to those of WaterGAP. However, in this context, it must also be noted that the data-driven model of this study aims at one target hydrological component, while the hydrological model is targeting multiple components in the hydrological system. Also, the hydrological model is built based on its

embedded causal model (equations). Therefore, it can be free from misleading results driven by confoundedness and redundancy of the predictors. Also, enough information of the predictors that cannot be measured can still be given to the model through the relationships of predictors in the causal model, so the model can be built with observed data that are relatively easily obtained. Yet, since many components in hydrological models are usually calibrated with a few sets of data (usually river discharge), quite high inaccuracy for single components can exist (e.g., deep percolation), despite their high performance (non-unique solutions). Moreover, processes as described in hydrological models are known to be lacking scale-related theories and not necessarily valid at larger scale (Nearing et al., 2021).

4. Conclusion

Formerly, groundwater recharge estimating models are largely dependent on models with relatively low flexibility such as process-based models or regression models with a fixed structure. However, with the help of XAI, the advantage of using a model with high flexibility is being kept, while the result of the model is still explainable. XAI can help quantify the importance and sensitivities of predictors as well as diagnose misleading results. Also, the lack of large data has been one of the main obstacles, but recently, there have been large inputs of data for groundwater recharge rate, collated from a few comprehensive meta-studies, and additional data for the Arabian Peninsula were added to the database in this study. Since the model is structured with high flexibility based on the quite large observed data set, inversely, there is now an opportunity to learn the process of target mechanism from the model. The data-driven neural network model for estimating groundwater recharge followed mostly the corresponding domain knowledge on the impact and interaction of predictors. Despite the absence of the groundwater recharge rate estimates in permafrost area, this might imply that the collated data have been sufficient to learn some governing mechanisms of groundwater recharge and achieve predictions with an accuracy comparable to process-based conventional hydrological models. Also, this large data of ground-based estimates for groundwater recharge collated in this study can be used to optimize or validate a hydrological model or further applied for data driven model applications.

However, it needs to be emphasized that, in any case, AI and XAI must be applied with a fair degree of caution and thoughtfulness. Despite the high performance, neural network models can also generate misleading results due to limitations in observational data and predictors such as statistical issues like data skewness, confoundedness and redundancy among predictors. At the same time, deficiencies in the data sets themselves might influence the reliability of predictions. This includes inconsistent spatial and temporal scales of the data as well as data quality. Despite a steady increase and improvement in global data sets, this is an aspect that needs to be considered especially for larger studies outside the relatively few well-monitored regions. Thus, it is recommended that the results of XAI are carefully interpreted using a causal model and validation, taking hydrological process understanding into account, especially when the predictions involve future projections, such as impacts of climate change.

Can XAI offer a new perspective for groundwater recharge estimation? Yes, we think so, especially considering that big data in hydrology have started to become an issue with the advent of high-resolution measuring methods. Successively, data-driven models and machine learning based methods are expected to be seen much more in this research field with big data. By applying XAI, we identify and quantify a large range of possible connections of extensive predictors and groundwater recharge process. In some cases, it's been possible to go beyond the usual measures as correlation. What AI model and XAI can offer to the domain is to point out the direction of further research, and analyzing all these connections is a large amount of work that we intend to do in the further studies.

Furthermore, data-driven models are not dependent on human perceptions that are built on one's experience and carry a certain risk of not being able to foresee all relevant processes—especially when dealing with large spatial scales (Nearing et al., 2021). In this context, learning from data can provide additional insights, although this also requires explanations through causal models. Recently, besides theories in hydrology, causal models are actively being studied to answer this question in the field of data science (Pearl & Mackenzie, 2018). Since the real causal relationships are not known, observational causal inference methods based on data analysis might be applied further to understand causality (Chernozhukov et al., 2018).

Data Availability Statement

The observational data of groundwater recharge rate on the Arabian Peninsula collected in this study is available at Zenodo (https://doi.org/10.5281/zenodo.10475566), and the other data sets used in this study are open sources and can be obtained from the references mentioned in Section 2.1.1 and Table 1.

Acknowledgments
The authors thank three anonymous reviewers and the editor Xavier Sánchez-Vila for their constructive reviews that helped us to improve the paper. Moreover, our gratitude is extended to all authors of individual studies on groundwater recharge estimates, whose works have been collated and utilized in this study. Open Access funding enabled and organized by Projekt DEAL.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2016). TensorFlow: Large-scale machine learning on heterogeneous distributed systems. https://doi.org/10.48550/arXiv.1603.04467

Adler, R. F., Gu, G., Wang, J., Huffman, G. J., Curtis, S., & Bolvin, D. (2008). Relationships between global precipitation and surface temperature on interannual and longer timescales (1979–2006). *Journal of Geophysical Research*, *113*(D22), 2008JD010536. https://doi.org/10.1029/2008JD010536

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., et al. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, *99*, 101805. https://doi.org/10.1016/j.inffus.2023.101805

Beck, H. E., Zimmermann, N. E., McVicar, T. R., Vergopolan, N., Berg, A., & Wood, E. F. (2018). Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Scientific Data*, *5*(1), 180214. https://doi.org/10.1038/sdata.2018.214

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), e2022GL099010. https://doi.org/10.1029/2022GL099010

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*(1–2), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Blöschl, G., Bierkens, M. F. P., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., et al. (2019). Twenty-three unsolved problems in hydrology (UPH) – A community perspective. *Hydrological Sciences Journal*, *64*(10), 1141–1158. https://doi.org/10.1080/02626667.2019.1620507

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

Carrera-Hernández, J. J., Mendoza, C. A., Devito, K. J., Petrone, R. M., & Smerdon, B. D. (2011). Effects of aspen harvesting on groundwater recharge and water table dynamics in a subhumid climate. *Water Resources Research*, *47*(5). https://doi.org/10.1029/2010WR009684

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68. https://doi.org/10.1111/ectj.12097

Chollet, F. (2017). *Deep learning with Python*. Manning Publications Company.

Crosbie, R. S., Jolly, I. D., Leaney, F. W., & Petheram, C. (2010). Can the dataset of field based recharge estimates in Australia be used to predict recharge in data-poor areas? *Hydrology and Earth System Sciences*, *14*(10), 2023–2038. https://doi.org/10.5194/hess-14-2023-2010

Cuthbert, M. O., Gleeson, T., Moosdorf, N., Befus, K. M., Schneider, A., Hartmann, J., & Lehner, B. (2019). Global patterns and dynamics of climate–groundwater interactions. *Nature Climate Change*, *9*(2), 137–141. https://doi.org/10.1038/s41558-018-0386-4

de Graaf, I. E. M., Sutanudjaja, E. H., van Beek, L. P. H., & Bierkens, M. F. P. (2015). A high-resolution global-scale groundwater model. *Hydrology and Earth System Sciences*, *19*(2), 823–837. https://doi.org/10.5194/hess-19-823-2015

Didan, K. (2021). MODIS/Terra vegetation indices monthly L3 global 0.05Deg CMG V061 [Dataset]. *NASA EOSDIS Land Processes DAAC*. https://doi.org/10.5067/MODIS/MOD13C2.061

Döll, P., & Fiedler, K. (2008). Global-scale modeling of groundwater recharge. *Hydrology and Earth System Sciences*, *12*(3), 863–885. https://doi.org/10.5194/hess-12-863-2008

Fan, Y., Miguez-Macho, G., Jobbágy, E. G., Jackson, R. B., & Otero-Casal, C. (2017). Hydrologic regulation of plant rooting depth. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(40), 10572–10577. https://doi.org/10.1073/pnas.1712381114

Gleeson, T., Wagener, T., Döll, P., Zipper, S. C., West, C., Wada, Y., et al. (2021). GMD perspective: The quest to improve the evaluation of groundwater representation in continental-to global-scale models. *Geoscientific Model Development*, *14*(12), 7545–7571. https://doi.org/10.5194/gmd-14-7545-2021

Global Soil Data Task Group. (2000). *Global gridded surfaces of selected soil characteristics (IGBP-DIS)*. ORNL Distributed Active Archive Center. https://doi.org/10.3334/ORNLDAAC/569

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational & Graphical Statistics*, *24*(1), 44–65. https://doi.org/10.1080/10618600.2014.907095

Haaf, E., Giese, M., Reimann, T., & Barthel, R. (2023). Data-driven estimation of groundwater level time-series at unmonitored sites using comparative regional analysis. *Water Resources Research*, *59*(7), e2022WR033470. https://doi.org/10.1029/2022WR033470

Harris, I., Osborn, T. J., Jones, P., & Lister, D. (2020). Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Scientific Data*, *7*(1), 109. https://doi.org/10.1038/s41597-020-0453-3

Hartmann, J., & Moosdorf, N. (2012). The new global lithological map database GLiM: A representation of rock properties at the Earth surface. *Geochemistry, Geophysics, Geosystems*, *13*(12). https://doi.org/10.1029/2012GC004370

Healy, R. W., & Scanlon, B. R. (2010). *Estimating groundwater recharge*. Cambridge University Press. https://doi.org/10.1017/CBO9780511780745

Her, Y., & Seong, C. (2018). Responses of hydrological model equifinality, uncertainty, and performance to multi-objective parameter calibration. *Journal of Hydroinformatics*, *20*(4), 864–885. https://doi.org/10.2166/hydro.2018.108

Ilstedt, U., Bargués Tobella, A., Bazié, H. R., Bayala, J., Verbeeten, E., Nyberg, G., et al. (2016). Intermediate tree cover can maximize groundwater recharge in the seasonally dry tropics. *Scientific Reports*, *6*(1), 21930. https://doi.org/10.1038/srep21930

Irrgang, C., Boers, N., Sonnewald, M., Barnes, E. A., Kadow, C., Staneva, J., & Saynisch-Wagner, J. (2021). Towards neural Earth system modelling by integrating artificial intelligence in Earth system science. *Nature Machine Intelligence*, *3*(8), 667–674. https://doi.org/10.1038/s42256-021-00374-3

Jaafarzadeh, M. S., Tahmasebipour, N., Haghizadeh, A., Pourghasemi, H. R., & Rouhani, H. (2021). Groundwater recharge potential zonation using an ensemble of machine learning and bivariate statistical models. *Scientific Reports*, *11*(1), 5587. https://doi.org/10.1038/s41598-021-85205-6

Jarvis, A., Guevara, E., Reuter, H. I., & Nelson, A. D. (2008). *Hole-filled SRTM for the Globe: Version 4: Data grid*. CGIAR Consortium for Spatial Information. Retrieved from http://srtm.csi.cgiar.org

Jasechko, S., Birks, S. J., Gleeson, T., Wada, Y., Fawcett, P. J., Sharp, Z. D., et al. (2014). The pronounced seasonality of global groundwater recharge. *Water Resources Research*, *50*(11), 8845–8867. https://doi.org/10.1002/2014WR015809

Jasechko, S., Seybold, H., Perrone, D., Fan, Y., Shamsudduha, M., Taylor, R. G., et al. (2024). Rapid groundwater decline and some cases of recovery in aquifers globally. *Nature*, *625*(7996), 715–721. https://doi.org/10.1038/s41586-023-06879-8

Jasechko, S., & Taylor, R. G. (2015). Intensive rainfall recharges tropical groundwaters. *Environmental Research Letters*, *10*(12), 124015. https://doi.org/10.1088/1748-9326/10/12/124015

Jung, H., Saynisch-Wagner, J., & Schulz, S. (2024). Estimates of the groundwater recharge rate on the Arabian Peninsula [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.10475566

Keese, K. E., Scanlon, B. R., & Reedy, R. C. (2005). Assessing controls on diffuse groundwater recharge using unsaturated flow modeling. *Water Resources Research*, *41*(6). https://doi.org/10.1029/2004WR003841

Kim, J. H., & Jackson, R. B. (2012). A global analysis of groundwater recharge for vegetation, climate, and soils. *Vadose Zone Journal*, *11*(1). https://doi.org/10.2136/vzj2011.0021RA

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Liu, F., Ting, K., & Zhou, Z. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data*, *6*(3), 1–39. https://doi.org/10.1145/2133360.2133363

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. https://doi.org/10.48550/arXiv.1705.07874

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), 034012. https://doi.org/10.1088/1748-9326/abd661

Markovich, K. H., Manning, A. H., Condon, L. E., & McIntosh, J. C. (2019). Mountain-block recharge: A review of current understanding. *Water Resources Research*, *55*(11), 8278–8304. https://doi.org/10.1029/2019WR025676

Martinsen, G., Bessiere, H., Caballero, Y., Koch, J., Collados-Lara, A. J., Mansour, M., et al. (2022). Developing a pan-European high-resolution groundwater recharge map – Combining satellite data and national survey data using machine learning. *The Science of the Total Environment*, *822*, 153464. https://doi.org/10.1016/j.scitotenv.2022.153464

Moeck, C., Grech-Cumbo, N., Podgorski, J., Bretzler, A., Gurdak, J. J., Berg, M., & Schirmer, M. (2020). A global-scale dataset of direct natural groundwater recharge rates: A review of variables, processes and relationships. *The Science of the Total Environment*, *717*, 137042. https://doi.org/10.1016/J.SCITOTENV.2020.137042

Mohan, C., Western, A. W., Wei, Y., & Saft, M. (2018). Predicting groundwater recharge for varying land cover and climate conditions – A global meta-study. *Hydrology and Earth System Sciences*, *22*(5), 2689–2703. https://doi.org/10.5194/hess-22-2689-2018

Morbidelli, R., Saltalippi, C., Flammini, A., & Govindaraju, R. S. (2018). Role of slope on infiltration: A review. *Journal of Hydrology*, *557*, 878–886. https://doi.org/10.1016/j.jhydrol.2018.01.019

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

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., et al. (2021). What role does hydrological science play in the age of machine learning? *Water Resources Research*, *57*(3), e2020WR028091. https://doi.org/10.1029/2020WR028091

Oki, T., & Kanae, S. (2006). Global hydrological cycles and world water resources. *Science*, *313*(5790), 1068–1072. https://doi.org/10.1126/science.1128845

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect* (2nd ed.). Basic Books.

Price, K., Jackson, C. R., & Parker, A. J. (2010). Variation of surficial soil hydraulic properties across land uses in the southern Blue Ridge Mountains, North Carolina, USA. *Journal of Hydrology*, *383*(3), 256–268. https://doi.org/10.1016/j.jhydrol.2009.12.041

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al. (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

Ross, C. W., Prihodko, L., Anchang, J., Kumar, S., Ji, W., & Hanan, N. P. (2018). Data descriptor: HYSOGs250m, global gridded hydrologic soil groups for curve-number-based runoff modeling background & summary. *Scientific Data*, *5*(1), 180091. https://doi.org/10.1038/sdata.2018.91

Scanlon, B. R., Keese, K. E., Flint, A. L., Flint, L. E., Gaye, C. B., Edmunds, W. M., & Simmers, I. (2006). Global synthesis of groundwater recharge in semiarid and arid regions. *Hydrological Processes*, *20*(15), 3335–3370. https://doi.org/10.1002/HYP.6335

Scanlon, B. R., Zhang, Z., Save, H., Sun, A. Y., Müller Schmied, H., van Beek, L. P. H., et al. (2018). Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data. *Proceedings of the National Academy of Sciences*, *115*(6), E1080–E1089. https://doi.org/10.1073/pnas.1704665115

Schulz, S., Re, V., Kebede, S., Abdalla, O., Wang, W., Simmons, C., & Michelsen, N. (2024). Preface: Hydrogeology of arid environments. *Hydrogeology Journal*, *32*(1), 1–8. https://doi.org/10.1007/s10040-023-02763-x

Seibert, J., Clerc-Schwarzenbach, F. M., & van Meerveld, H. J. (2024). Getting your money's worth: Testing the value of data for hydrological model calibration. *Hydrological Processes*, *38*(2), e15094. https://doi.org/10.1002/hyp.15094

Shamsudduha, M., Taylor, R. G., Ahmed, K. M., & Zahid, A. (2011). The impact of intensive groundwater abstraction on recharge to a shallow regional aquifer system: Evidence from Bangladesh. *Hydrogeology Journal*, *19*(4), 901–916. https://doi.org/10.1007/s10040-011-0723-4

Soltani, S. S., Ataie-Ashtiani, B., & Simmons, C. T. (2021). Review of assimilating GRACE terrestrial water storage data into hydrological models: Advances, challenges and opportunities. *Earth-Science Reviews*, *213*, 103487. https://doi.org/10.1016/J.EARSCIREV.2020.103487

Verdin, K. L., & Survey, U. S. G. (2017). Hydrologic derivatives for modeling and analysis—A new global high-resolution database. In *Data Series*. https://doi.org/10.3133/ds1053

Vereecken, H., Weihermüller, L., Assouline, S., Šimůnek, J., Verhoef, A., Herbst, M., et al. (2019). Infiltration from the Pedon to global grid scales: An overview and Outlook for land surface modeling. *Vadose Zone Journal*, *18*(1). https://doi.org/10.2136/vzj2018.10.0191

Whitford, W. G., & Duval, B. D. (2020). Chapter 3 - Characterization of desert climates. In W. G. Whitford & B. D. Duval (Eds.), *Ecology of desert systems* (2nd ed., pp. 47–72). Academic Press. https://doi.org/10.1016/B978-0-12-815055-9.00003-5

Wunsch, A., Liesch, T., & Broda, S. (2022). Deep learning shows declining groundwater levels in Germany until 2100 due to climate change. *Nature Communications*, *13*(1), 1221. https://doi.org/10.1038/s41467-022-28770-2

Yenehun, A., Dessie, M., Nigate, F., Belay, A. S., Azeze, M., Camp, M. V., et al. (2022). Spatial and temporal simulation of groundwater recharge and cross-validation with point estimations in volcanic aquifers with variable topography. *Journal of Hydrology: Regional Studies*, *42*, 101142. https://doi.org/10.1016/J.EJRH.2022.101142