water

MDPI

Review

A Review on Interpretable and Explainable Artificial Intelligence in Hydroclimatic Applications

Hakan Başağaoğlu [1,*], Debaditya Chakraborty [2,*], Cesar Do Lago [2], Lilianna Gutierrez [3], Mehmet Arif Şahinli [4], Marcio Giacomoni [2], Chad Furl [1], Ali Mirchi [5], Daniel Moriasi [6] and Sema Sevinç Şengör [7]

[1] Edwards Aquifer Authority, San Antonio, TX 78215, USA; cfurl@edwardsaquifer.org
[2] School of Civil and Environmental Engineering and Construction Management, University of Texas at San Antonio, San Antonio, TX 78207, USA; cesar.dolago@utsa.edu (C.D.L.); marcio.giacomoni@utsa.edu (M.G.)
[3] Department of Biomedical and Chemical Engineering, University of Texas at San Antonio, San Antonio, TX 78207, USA; lilianna.gutierrez@utsa.edu
[4] Department of Agricultural Economics, Faculty of Agriculture, Ankara University, Ankara 06110, Turkey; asahinli@ankara.edu.tr
[5] Department of Biosystems & Agricultural Engineering, Oklahoma State University, Stillwater, OK 74078, USA; amirchi@okstate.edu
[6] USDA-ARS Grazinglands Research Laboratory, El Reno, OK 73036, USA; daniel.moriasi@usda.gov
[7] Department of Environmental Engineering, Middle East Technical University, Ankara 06800, Turkey; ssengor@metu.edu.tr
[*] Correspondence: hbasagaoglu@edwardsaquifer.org (H.B.); debaditya.chakraborty@utsa.edu (D.C.)

Abstract: This review focuses on the use of Interpretable Artificial Intelligence (IAI) and eXplainable Artificial Intelligence (XAI) models for data imputations and numerical or categorical hydroclimatic predictions from nonlinearly combined multidimensional predictors. The AI models considered in this paper involve Extreme Gradient Boosting, Light Gradient Boosting, Categorical Boosting, Extremely Randomized Trees, and Random Forest. These AI models can transform into XAI models when they are coupled with the explanatory methods such as the Shapley additive explanations and local interpretable model-agnostic explanations. The review highlights that the IAI models are capable of unveiling the rationale behind the predictions while XAI models are capable of discovering new knowledge and justifying AI-based results, which are critical for enhanced accountability of AI-driven predictions. The review also elaborates the importance of domain knowledge and interventional IAI modeling, potential advantages and disadvantages of hybrid IAI and non-IAI predictive modeling, unequivocal importance of balanced data in categorical decisions, and the choice and performance of IAI versus physics-based modeling. The review concludes with a proposed XAI framework to enhance the interpretability and explainability of AI models for hydroclimatic applications.

Keywords: explainable artificial intelligence; multidimensional data; nonlinearity; explanatory methods; hydroclimatic applications

check for updates

Citation: Başağaoğlu, H.; Chakraborty, D.; Lago, C.D.; Gutierrez, L.; Şahinli, M.A.; Giacomoni, M.; Furl, C.; Mirchi, A.; Moriasi, D.; Şengör, S.S. A Review on Interpretable and Explainable Artificial Intelligence in Hydroclimatic Applications. *Water* **2022**, *14*, 1230. https://doi.org/10.3390/w14081230

Academic Editor: Ana Iglesias

Received: 20 February 2022
Accepted: 7 April 2022
Published: 11 April 2022

1. Introduction

Recent advancements in sensors, extended measurement networks, increasing use of remote sensing products, improvements in accuracy and reliability of monitoring devices with more frequent automated data acquisition capabilities, and enhanced storage and communication technologies are generating unprecedented volumes of high dimensional hydroclimatic data more than ever before [1–3]. At the same time, Artificial Intelligence (AI) algorithms have emerged as versatile tools to unfold data-driven novel information from the sheer volume of multidimensional data combined in nonlinear and highly interactive ways, where such analyses were previously unimaginable using conventional time-series or simple statistical techniques [4].

In this review, we focus on interpretable AI (IAI) and explainable AI (XAI) models for supervised regression or categorical predictions in hydroclimatic domains. 'Interpretability' here refers to the ability of the AI models to unveil the nonlinear correlative effects between the predictors and predictands to a degree that humans can understand the rationale behind the predictions [5,6]. 'Explainability' here refers to a collection of interpretations from IAI models with further contextual information stemming from domain knowledge and related analysis [7], which are used to justify decisions, enhance control, improve models, and discover new knowledge [8]. The ability to understand the overall predictive behavior by ranking predictors with respect to their importance in predictions and construction of testable hypotheses to unveil critical conditions for the predicted conditions to occur probabilistically are examples of explanability measures. Through the explanatory measures, the users can peek into the internal logic and mechanics of an AI system. Thus, interpretability is the prerequisite for explainability, and the explainability is essential for the scientific value of the outcome [7].

The taxonomy of AI and the screening process used to select the papers for our review is shown in Figure 1. The review specifically focuses on the interpretability and explainability of tree-based ensemble AI models, based on the bagging and boosting algorithms, which have been successfully implemented in recent years for data imputations, inferences, and predictions in diverse hydroclimatic applications. A recent survey indicated that tree-based model structures have been used as a base learner in ensemble AI models in 42% of the models in hydrologic applications [9]. Although tree-based ensemble models were considered as a black-box model by some scholars [10], we argue that these models are amenable to be coupled with the explanatory methods such as SHaply Additive eXplanation (SHAP) [11,12] and Local Interpretable Model-agnostic Explanations (LIME) [13] to achieve enhanced interpretability and explanability of the AI-based predictions in diverse domains, and hence, they are indeed not black-box models, as demonstrated in Refs. [14–23].

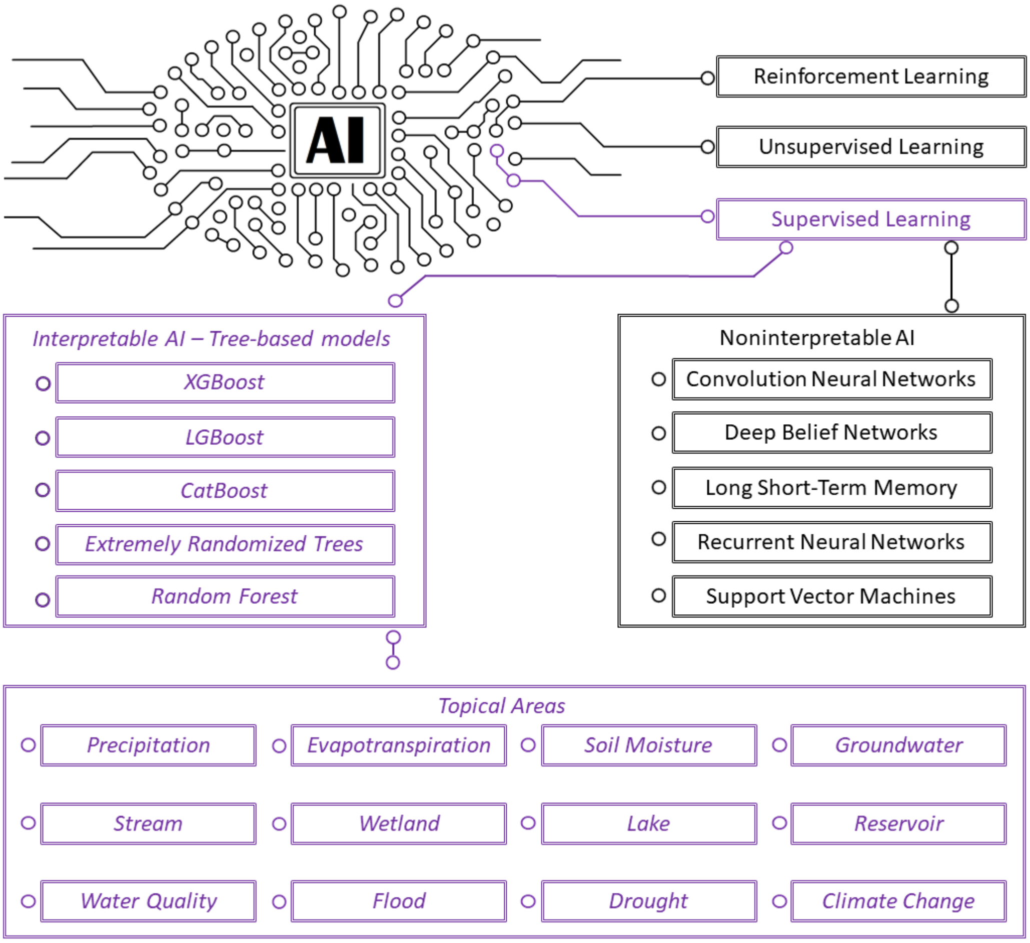

Figure 1. Taxonomy of AI and topical areas selected for this study. This study focuses on subject matters that are embedded within the purple boxes.

We focus on the following questions in reviewing recent IAI- and XAI-based analysis in diverse hydroclimatic problems at different spatiotemporal scales, although not all questions were answered in each reviewed paper:

- Which predictors and predictands are used in the IAI-based analysis? What is the size, type (e.g., static or time-variant), and sampling interval of multidimensional input data? Are the chosen predictors representative of the underlying physics of the problem tackled? Is the use of surrogate variables in IAI-based analyses acceptable for the regions with scarce data?
- Are the explanatory methods properly and effectively coupled with the AI models (leading to XAI models) to assess the importance of the predictors in predictions, explain the interdependencies and interrelations between the predictors in estimating the predictands, justify the IAI-based decisions, and explore new knowledge?
- Are class imbalances properly addressed in categorical IAI and XAI modeling applications?
- Under what conditions could in-depth domain knowledge become critical? Would domain knowledge allow flexibility for the choice of predictive variables in IAI/XAI-based analyses?
- Should multiple IAI/XAI models be used independently or should the results from multiple IAI/XAI models be a weighted-average? Can IAI/XAI and non-IAI models be used in a hybrid form to enhance prediction accuracy? How do prediction performances of IAI/XAI and non-IAI models compare in different domains with different data types and sizes?
- Are there any attempts toward interventional XAI modeling in hydroclimatic applications to relax the nonstationary assumption in AI-based analyses?
- How do IAI/XAI models perform against physics-based models? Are there applications, in which IAI/XAI models fail to provide reliable results?

The paper is structured as follows: In Section 2, we provide the definition of IAI and XAI models, and a brief discussion on the use of explanatory methods to transform AI models to IAI and XAI models. In Section 3, we describe the tree-based ensemble IAI models considered in this paper. Section 4 provides a review of recent studies on the use of IAI and XAI models for data imputations, inferences, and predictions. The review focuses on IAI and XAI models-based analyses in hydroclimatic applications for enhanced inference and prediction of climatic features (evapotranspiration, precipitation), subsurface features (soil moisture, groundwater potential and levels), surface water features (streamflow, water levels in wetlands, lakes, and reservoirs), water quality features (water quality in surface waters and aquifers), extreme climate events (flood hazard and drought risks), and climate change impacts on the hydrological cycle.

2. IAI and XAI Models

There is no concrete mathematical definition, formality, or measured metric for interpretability or explainability [8,24]. It was proposed that interpretability should be split into two broad categories: the first one is related to transparency, seeking an answer to 'how does the model work?', and the second one is related to post-hoc explanations, seeking an answer to 'what else can the model tell?' [25]. Some researchers argued that the notion of interpretability often depends on the domain of application [26], therefore it cannot be fitted into a tight definition. While interpretability was equated to explainability by some researchers [6], interpretability was considered to be a broader term than explainability by others [24]. In this review, the AI model is deemed to be interpretable if it is capable of unveiling the rationale behind the predictions that is understandable by a human [6] and is deemed to be explainable if it is capable of justifying the decision made, enhancing control, improving the decision, and revealing new knowledge [8] from a collection of interpretations from an IAI model coupled with contextual information [7].

Although tree-based ensemble models interfaced with the explanatory methods (e.g., SHAP) are interpretable and explainable [18], many accurate decision support systems (e.g., Deep Learning (DL) models) have been constructed as black boxes, in which the systems hide their internal logic from the user [27]. This makes it harder to assign physical meaning and interpretation to the features estimated by the AI model [28]. Conversely, the explainability of AI-based decisions is linked to trust and user behaviors. Explanations of why certain decisions are made help to build trust with users. Such trust is formed based on the extent the users understand the explanations. XAI models help users understand how forecasts arise and how they can be influenced or adjusted to arrive at workable predictions [29], thereby bringing fairness and accountability into the AI-based decision-making process [30]. Conversely, lack of explanation constitutes both a practical and an ethical issue in regards to accountability and trustworthiness of the results, 'opaque' decisions, and risks for inadvertently making the wrong decisions [10]. Required levels of explanations are often dependent on the main objectives of the application. According to the granularity of the analysis, strategies for XAI decisions typically focus on local (understanding a single prediction) and global (understanding the entire model behavior) explanations. Some researchers noted that omitting explainability in AI-based clinical decision support systems poses a threat to core ethical values in medicine and may have detrimental consequences for individual and public health [31]. This view was contended by other researchers who argued that opaque decisions are common in medicine, where explainability of the results would be less important than accuracy of the result if the accuracy is verified indirectly or empirically [32]. We expect to see similar conflicting arguments for the use of XAI models in hydroclimatic predictions in the near future.

The explanatory methods (e.g., SHaply Additive exPlanation (SHAP) [11,12] and Local Interpretable Model-agnostic Explanations (LIME) [13]) have been implemented only in a handful of hydroclimatic problems to date [17,18,20,33–35]. However, these methods have been used in diverse domains to enhance the explainability of AI-based decisions by unveiling the dependencies between the predictors and predictands, reducing the dimensionality of the input space, identifying the inflection points above or below which the predictands respond negatively or positively to the changes in the values of the predictors, and setting up testable hypotheses to unveil critical conditions for the predicted conditions to occur probabilistically [19,21,36,37].

3. Tree-Based Ensemble IAI Models Considered in This Review

Tree-based ensemble algorithms combine multiple simple decision tree models trained by the same learning algorithms and use bagging, boosting, and stacking algorithms to reduce variance and deviation [38]. In this review, we mainly focus on the implementation of Extreme Gradient Boosting (XGBoost) [39], Light Gradient Boosting (LGBoost) [40], Categorical Boosting (CatBoost) [41], Extremely Randomized Trees (ERT) [42], and Random Forest (RF) [43] models in hydroclimatic applications. RF and ERT are bagging-based algorithms while XGBoost, LGBoost, and CatBoost are boosting-based algorithms. In the bagging algorithms, when the decision trees are built, the decision trees run in parallel independently and do not interact with each other. In building decision trees, RF subsamples the input data with replacement using the bootstrap method, whereas ERT uses the entire original sample. Although RF chooses the local optimal split, the ERT chooses the split randomly in making decisions. Once the split is chosen, both algorithms choose the best one among all the subset of features. Gradient boosted trees are an ensemble of weak classifiers or regressors (e.g., a decision tree model) where multiple weaker models are combined to produce a stronger model. In the boosting algorithm, multiple trees are grown sequentially using the information from the existing trees. A new decision tree is generated by improving the performance of the tree generated in the previous iteration. Although XGBoost splits the trees depth-wise or level-wise, LGBoost splits the trees leafwise. These AI models are amenable to be fused with the explanatory methods, such as SHAP or LIME, to form IAI and XAI models that are capable of providing enhanced interpretability and explainability in AI-based decisions. Therefore, in light of the definitions in Section 2, these AI models are indeed IAI models, which can be upgraded to XAI models if they can justify the AI-based decisions and reveal new knowledge with the help of explanatory methods.

4. IAI and XAI Applications in Hydroclimatic Domains

The review focuses on recent predictive IAI and XAI models published in and after 2018, except for a few noteworthy papers published earlier. The review considers IAI-based supervised predictive analysis based on multidimensional nonlinearly related numerical or categorical data. In these applications, the input dataset, including predictors and observed predictands, is split into training and testing datasets typically at the ratio of 70:30 to 90:10. After the IAI model is trained with the training data, its prediction accuracy is verified using the test data unseen by the IAI model during model training. If the predicted target variables computed from the predictors in the test data statistically agree with the observed target variables in the test data (based on, for example, the coefficient of determination, root-mean-square error), the IAI model can then be used as a predictive tool. The IAI models are often optimized using grid search hyperparameter tuning to search for the global optimal solution over a nonlinear solution space. In the applications discussed in the subsequent sections, IAI and/or XAI models were used to enhance the interpretability and explainability of the AI-based decisions.

For the categorical supervised IAI and XAI models, the confusion matrix or area under receiver operating characteristics curve (AUROC) is commonly used to assess the prediction accuracy of the models. For a two-class classification problem, the confusion matrix reports true positives, true negatives, false positives, and false negatives. The AUROC is constructed based on true positive and false positive rates. Prediction accuracy of the model, however, is sensible when balanced classification data is used in model training and testing.

4.1. Data Imputations Using IAI Models

Imputation is a process that replaces the missing data by reasonable values, and AI algorithms have shown to handle missing data efficiently and accurately [18] while avoiding assumptions about the statistical distribution of the data. The missing data can be categorized as: (i) missing completely at random, in which the probability of missing sample is independent of the observed and unobserved data; (ii) missing at random, in which the incomplete data differ from the complete data, but the pattern of missingness is predictable from the remaining dataset; and (iii) nonrandom missingness, where the pattern of data is nonnegligible and is not predictable from the rest of the data [44]. Although random or sequential missing data, as described above in (ii), are common in hydrological models [**?**] and occur when no data value is stored during observation, large volume and long stretches of missing data, outliers, or erroneously entered data are serious problems in data quality and mining, which could adversely impact the AI-based prediction and decision-making process.

Precipitation (P) is a discontinuous hydroclimatic variable, especially in arid and semi-arid regions. Conventional interpolation methods used for data imputation often overestimate the number of rainy days and underestimate the extreme precipitation events, and hence, do not preserve the probability distribution of P [46,47]. The RF model was used to impute 64% of the missing daily P data over 15,219 days of the sampling period across a network of 112 rain gauges covering an area of around 3000 km^2 in Spain [48]. Use of the RF model also involved a binary categorical prediction for each day in the test data whether the day was to be labeled as 'rain day' or 'no-rain day'. Subsequently, the RF model predicted daily P totals only for the days predicted to be rainy. In our opinion, using another AI-based classifier to determine whether a specific day was rainy or non-rainy was unnecessary, as the RF model can readily predict a time series of daily rainfall totals. Nonetheless, the RF model was found to be more efficient for imputing random missing data than sequentially missing data, but overestimated daily P as the number of rainy days were over-predicted.

Multiple data types could have missing values at different lengths and frequencies at some sites. AI-based 'sequential transfer learning' method was developed to impute long stretches of missing data in multiple climate variables at one of the meteorological stations (labeled as SGD) in a semi-dry region in Texas, USA [18]. Twenty percent of the entire climate dataset, including ∼336 days of consecutive measurements, in addition to 10% solar radiation (R_s) data were missing. In the imputation method, the XGBoost model was trained to learn the dynamic relationship between the non-missing air temperature (T_a) data at the SGD and two other meteorological stations with nearly complete data. The trained model was then used to predict missing T_a at SGD. Similar steps were taken to impute the missing atmospheric pressure (P_a) in the SGD dataset. Next, the XGBoost model was trained to learn the dynamic relationship between the non-missing relative humidity (RH), T_a and P_a at SGD. Using the trained model, the missing RH were predicted from T_a and P_a that were predicted in the previous steps. Then, the authors modeled the non-missing R_s with respect to T_a, P_a, and RH at SGD, and subsequently predicted the missing R_s using the predicted T_a, RH, and P_a. They modeled the non-missing wind velocity (u_2) from T_a, P_a, RH, R_s at SGD, and then predicted missing U_w using the predicted T_a, P_a, RH, and R_s. In the end, the imputed data produced monthly reference crop evapotranspiration (ET_o) values that were trend-wise and magnitude-wise in agreement with the predicted ET_o at the neighboring stations with more complete data.

Streamflow (Q_s) is a continuous variable for perennial streams, but a discontinuous variable for ephemeral streams. Flow rates could be affected by natural hydroclimatic processes under climate change, in addition to alterations by human activities, such as runoff in urbanized regions, diversions for irrigation, dam construction, hydropower generation, and changes in watershed characteristics, which could complicate data imputations. The MissForest algorithm [49], extended from the RF model was used to impute daily Q_s time series at 122 gauges with the missing data <50% of the time from 1970 to 2016 in data-scarce regions over multiple climate zones in Chile [50]. The RF model was trained using Q_s data from gauges with the least missing data, and then the trained model was used to infill missing Q_s data at other gauges with more missing data. We call their approach 'transfer learning'. The authors infilled the missing Q_s data used for model training by setting them to the average flow rates at that particular gauge. This approach, however, could introduce errors, if the missing values—which were set to the average value—in the training data occurred during extreme events. Additional errors could be introduced, if the hydrologic characteristics of the streams used for model training and that of targeted streams with missing Q_s data are not commensurable. For example, the RF model trained by the meandering stream Q_s data would not accurately represent the discharge-stage rating curve for faster flowing streams, and hence, may not provide reliable infilled values for missing Q_s data. Nevertheless, using the imputed data, the authors reported that the predictive performance of the MissForest algorithm for infilling missing values did not change significantly for single missing data points or missing contiguous data points up to 60 days. They also reported that their transfer learning approach yielded satisfactory-good performance in imputing data for streams with natural flows, but the prediction performance decreased for the streams with altered flow conditions by man-made structures, and failed at the extreme case of hydropeaking. An important takeaway is that the IAI models could still achieve accurate data imputation for missing Q_s data in altered hydrologic conditions, which has been frequently encountered in practice.

4.2. Hydroclimatic Predictions Using IAI and XAI Models

This section discusses IAI- and XAI-based predictions in diverse nonlinear hydroclimatic processes from multidimensional predictors. A list of predictors and notations used for each hydroclimatic application is summarized in a table in the same section. Only repetitively used variables and notations, in addition to acronyms for the IAI and XAI models are provided in the Abbreviation section.

4.2.1. Evapotranspiration Predictions

Evapotranspiration (ET) is a critical indicator of global climate change [51], and its reliable prediction is imperative for irrigation, agriculture, and surface water and groundwater management and planning [52]. ET is the sum of evaporation from soil and transpiration from vegetation. It is often reported as the reference crop evapotranspiration (ET_o), actual evapotranspiration (ET_a), potential evapotranspiration from wet surfaces (ET_p), or surfaces covered by large volume of water, such as wetlands or lakes (E_{sw}). ET_o is commonly predicted from a time series of meteorological predictors, including R_s, RH, T_a, U_w, and P_a. ET from open water surfaces also include the surface water temperature (T_{sw}) as a predictor. Terrestrial ET_a requires information on vegetation cover.

The FAO56-Penman Monteith Equation (PME) [53] has been commonly used to predict ET_o from R_s, T_a, RH, U_w, and P_a; however, complete meteorologic data are not available at some locations across the globe. Therefore, AI-based ET_o predictions from incomplete meteorological data have been examined in the literature. Performance of several tree-based, kernel-based, and curve-based AI models were compared in predicting daily ET_o from daily minimum and maximum T_a ($T_{a,min}$, $T_{a,max}$), and P at 14 stations from 2001 to 2015 in different climate zones in China [54]. Use of P as a predictor for ET_o, however, is uncommon and inconsistent with the PME. The authors assumed that P would represent RH especially in (sub)tropical-humid regions, which remains questionable. Based on this assumption and using 70% of the data to train the models, the authors concluded that Support Vector Machine (SVM) predicted daily ET_o with the highest accuracy while outperforming the XGBoost model. In a different study, prediction accuracy of the CatBoost, RF, and Generalized Regression Neural Network models were compared in estimating ET_o in arid and semi-arid regions in China [55]. Eight different combinations of $T_{a,min}$, $T_{a,max}$, U_w, RH, and R_s that were monitored from 1996 to 2015 at 15 stations were used as predictors. The 1996–2009 records were used for training and the 2010–2015 for testing the AI models. All the AI models performed well with incomplete data when only RH was not included as a predictor. Therefore, the authors recommended these AI models to predict ET_o at the sites with missing meteorological data. Conversely, CatBoost exhibited the best performance for all the combinations of data, and hence, was recommended for regions with similar climates. In a similar study, the performance of the optimized CatBoost, RF, and SVM models were compared in predicting daily ET_o at 12 weather stations in a subtropical region in China using different combinations of daily local meteorologic predictors of R_s, RH, $T_{a,min}$, $T_{a,max}$, and U_w under presumably water-scarce conditions [56]. The data from 2001–2010 and 2011–2015 were used to train and test the AI models, respectively. The authors concluded that all three AI models achieved satisfactory accuracy for ET_o prediction using either (R_s, $T_{a,min}$, $T_{a,max}$) or (U_w, RH, $T_{a,min}$,$T_{a,max}$), suggesting that either reduced predictors set can be used in water-scarce subtropical regions to predict ET_o. They also noted that when R_s, RH, $T_{a,min}$,$T_{a,max}$, and U_w were available, CatBoost yielded the best prediction accuracy. Conversely, SVM yielded the best prediction accuracy when some of the climatic data types were missing. In brief, [55,56] indicated that SVM (a non-IAI model) is a better predictor tool when some meteorologic data are missing, whereas CatBoost (an IAI model) is a better predictor tool when complete meteorologic data is available for AI-based ET_o predictions.

To further evaluate the relative performance of the IAI and non-IAI models in predicting ET_o from a complete set of multi-dimensional meteorological data, predictive accuracy of three optimized IAI models (XGBoost, RF, Linear Regression (LR)) and three optimized non-IAI models (DL, SVM, Long short-term memory (LSTM)) were compared in estimating daily ET_o computed by FAO56-PME from structured tabular data, including R_s, RH, T_a, U_w, and P_a over 4–5 years from multiple meteorological stations in a semi-arid region in Texas, USA [18]. Using 90% of the data for model training, prediction accuracy of the AI models was in the order of DL∼XGBoost>RF>LR∼SVM>LSTM. The authors concluded that the top-performing IAI model (XGBoost) exhibited comparable performance to the top performing non-IAI model (DL) in predicting daily ET_o. They developed a XAI model by coupling XGBoost with the SHAP method. The global SHAP analysis unveiled that the relative importance of the meteorological variables in ET_o prediction was in the order of $R_s > T_a > RH > U_w > P_a$ for the study area. Local SHAP and LIME analyses identified the inflection point of each predictor above or below which ET_o would increase. The inflection points were subsequently used to set up testable hypotheses using conditional probabilities to justify the XAI predictions and seek new knowledge. Considering the median observed ET_o value as the threshold, the authors showed that $R_s \leq 17.16$ kW/m^2, $T_a \leq 20.62$ °C, and $RH > 72.17\%$ at one of the sites, then ET_o would almost surely be below the median ET_o. To our knowledge, the XGBoost-SHAP-LIME model in this study was the first XAI model accompanied with testable hypotheses used for enhanced interpretability and explainability of daily ET_o predictions. The authors concluded that the XGBoost-based XAI framework displayed comparable performance to DL in predicting ET_o while holding physical interpretability of the predictors–predictand dynamics and unveiling the order of importance of the predictors in ET_o predictions. Unlike in [18], the feasibility of ET_o predictions from a single meteorological variable was investigated using the optimized XGBoost, RF, and Deep Neural Network (DNN) models by comparing the results against daily ET_o estimates from 32 years of local meteorological data in California, USA, including R_s, RH_{min}, RH_{max}, $T_{a,min}$, $T_{a,max}$, T_{avg} and U_w in [34]. Through the global Shapley and Gini-based feature importance analyses implemented with the RF and XGBoost models (led to XAI models), they concluded that R_s was the most influential predictor at three sites with different climatic conditions, in agreement with the conclusions in [18]. Using daily R_s as the sole predictor, daily FAO56-PME-computed ET_o as the predictand, and assigning 80% of the data to train the AI models, they concluded that DNN exhibited better prediction accuracy than XGBoost and RF. Their approach is different in the sense that it coupled the enhanced interpretability of the tree-based modeling and the high prediction capability of a noninterpretable DNN modeling for ET_o predictions.

A critical challenge with the earlier AI models was that the nonlinear relationship between climatic variables and the ET makes it difficult to account for inherent uncertainties [57]. This challenge was addressed in [17] by formulating a novel probabilistic IAI model, built on the hybrid XGBoost-NGBoost framework, to predict daily ET_o, ET_a, and E_{sw} using 3–5 years of daily meteorological data, including T_a, P_a, R_s, RH, U_w, month, T_{sw} (for E_{sw} prediction), and ET_o (for ET_a prediction) in south-central Texas, USA. Different from the earlier AI models, the hybrid XGBoost-NGBoost was able to produce not only point predictions, but also the probability distribution over the entire outcome space to quantify uncertainties associated with ET predictions. Using 90% of the data for model training, they demonstrated that probabilistic approach exhibited great potential to overcome data uncertainties, in which 100% of the ET_o, 89.9% of the E_{sw}, and 93% of the ET_a test data at three watersheds were within the models' 95% prediction intervals. Using the XGBoost-SHAP (a XAI model) analysis, the authors identified the top three influential features to be R_a, T_a, and RH for ET_o; T_{sw}, RH, and month for E_{sw}; and R_s, month, and RH for ET_a predictions at the semi-arid site.

4.2.2. Precipitation Predictions

The spatiotemporal variability and uncertainties in precipitation (P) measurements [58] make it a difficult hydroclimatic variable to work with, although it is a critical predictor for diverse hydroclimatic processes, such as surface runoff, flood, droughts, and aquifer recharge.

Stable isotopes of hydrogen and oxygen (δ^2H and $\delta^{18}O$) have been used as natural tracers to improve our understanding of hydrological and meteorological processes, including precipitation formation mechanisms [59]. An XGBoost model was recently used to explore interannual and longterm variability in monthly δ^2H and $\delta^{18}O$ time series of P using location and climate data [60]. The location data included the latitude (Lat), longitude (Lon), and altitude (AL) of the data site. The climate data included local climate data (e.g., P, T_a, R_s, U_s, vapor pressure (V_p)) and climate indices associated with large-scale atmospheric circulation (e.g., North Atlantic Oscillation index, the Scandinavian pattern) from a large number of gridded and time-series European data sources. In addition to the location and climate data, the month and season of the year and Köppen climate regions were used as predictors in the IAI model. The authors used 32,191 monthly observations of at least 1 stable isotope value from 270 stations for the period from 1960 to 2018, in which ∼20% of the data was used for model testing. They developed three independent IAI models using XGBoost, each for $\delta^2 H$, $\delta^{18} O$, and deuterium-excess (d-excess). They implemented three modeling steps: First, they ran these IAI models with the complete set of predictors. Next, they ran the models with the most important predictors only. Finally, each IAI model with the reduced predictors list also used the predicted predictands from the other two IAI models as the additional predictors. The overall IAI model was named Piso-AI, which is suitable to produce point and gridded monthly $\delta^2 H$ and $\delta^{18} O$ of P on demand, as the predictors are regularly updated. The model is useful to provide isotope input variables for ecological and hydrological application and paleoclimate proxy calibration. Prediction accuracy of the Piso-AI was reported to be better than the other predictive tools when the interannual variations were important. In our opinion, when/if the Piso-AI model is coupled with the explanatory methods such as SHAP, it could provide enhanced insights into predictors-predictands dynamics and overall results.

In hydroclimatic applications, gridded P data at coarser spatial scales can be used for local P estimates after downscaling, if they are shown to be representative of local climatic conditions. A RF model was used to assess the similarity of gridded monthly P and T_a data from external sources and locally observed data [61]. The suitability of seven external gridded P and five gridded T_a datasets with the spatial resolution of 0.25–0.50° was evaluated and ranked with respect to monthly observed local time-series data at 57 stations in Egypt for the period 1979–2014. Four grid points surrounding a station were interpolated to the station location using a inverse distance weighting method to generate time series of observed and gridded external data pairs at each station. The similarity index was defined as the number of times that the observed and gridded data at the particular station took the same path and placed in the same terminal node of the same tree in the RF model. Different from other IAI model applications, the entire data were used to train the IAI model. Using the RF model, the authors identified the most representative external climate datasets that agree with monthly local P and T_a at each station as well as their spatial variations. Because P influences many hydroclimatic processes and decisions, such IAI-based similarity assessments between remotely-sensed data with local measurements are indispensable in the development of local or regional water management decisions, especially when locally-measured P datasets are scarce or precarious.

The effect of P zoning on the accuracy of IAI-based downscaling of gridded P data from remote sensing precipitation products with a spatial resolution of 0.25° to ground-based P data with a spatial resolution of 1-km was investigated in [62]. Such P zoning was implemented to identify the predominant regional patterns of P variability. The study was conducted across the Lancang–Mekong River basin, which has a total area of about 795,000 km^2 covering parts of Southwest China, Myanmar, Laos, Cambodia, Thailand, and Vietnam and spans over multiple climate zones. The monthly satellite-based P data and ground-based P data from 29 meteorological stations and 261 rain gauge stations from 2000 to 2014 were used in the IAI analysis. Twelve meteorological stations and 229 rain gauges in 2001 (wet year), 200 rain gauges in 2005 (normal year), and 24 rain gauges in 2009 (dry year) were used for model validation. The authors used the iterative rotated empirical orthogonal function analysis of ground- and satellite-based P observations to delineate 6–7 P zones. They considered two cases: the first one did not involve discrete P zones and RF was used for the entire study area; in the second case, the study area was divided into different P zones and RF was applied independently to each zone. The authors implemented the RF model for downscaling, in which the latitude (Lat), longitude (Lon), altitude/elevation (AL), slope (SL), and normalized difference vegetation impacts ($NDVI$) were used as predictors and satellite-based P was used as the predictand. RF was trained and validated over the 0.25°- resolution data (coarser resolution). The validated RF was then used with a 1 km resolution data (Lat, Lon, AL, SL) to predict P at a 1 km resolution (finer resolution). The author concluded that zoning-based downscaling outperformed non-zoning-based downscaling in terms of the prediction accuracy. A permutation test implemented to assess the importance measure of the predictors revealed different importance rankings of the predictors responsible for P distributions at different spatial scales (e.g., at each P zone scale vs. at the entire study area scale). Thus, the spatial scale dependency of the predictors–predictand relation in this case could raise concern about the suitability of the RF model to predict P at finer resolutions, after being trained with data at coarser spatial resolutions without implementing proper scale-dependent error corrections, as discussed in [35].

4.2.3. Soil Moisture Predictions

Soil moisture (SM) is a spatially heterogeneous variable that affects surface runoff, base flow, aquifer recharge, and vegetation cover [63], and hence, it is a critical measure in hydrologic modeling and water and irrigation management decisions. Although remote sensing data have been commonly used to derive local-scale SM data, a mismatch between them is a challenge to overcome. Predictors used for IAI-based SM predictions in recent studies are summarized in Table 1.

Table 1. Factors and predictors used in IAI-based soil moisture predictions.

Factors	Predictors
Meteorologic	Precipitation (P), Temperature (T_a), Solar radiation (R_s), Wind speed (U_w), Relative humidity (RH), Sun hours (SH)
Hydro-climatic	Evapotranspiration (ET_o)
Topographic	Digital elevation model/Elevation (DEM), Slope (SL), Northness (Nt)
Land Surface	Land surface temperature (LST), Surface albedo (ALB)
Soil	Topographic wetness index (TWI), Column-average soil texture (ST)
Vegetation/Biophysical	Normalized difference vegetation index ($NDVI$), Surface albedo (ALB), Leaf area index (LAI), Crop type (CT), Location with respect to the canopy ($LCON$)

Similar to P data, gridded data at coarser-spatial scales are commonly used to predict local-scale SM. The RF model was used to downscale SM data (at tens of km-scale) from passive microwave surface SM products, including the soil moisture active passive (SMAP) and soil moisture and ocean salinity satellite (SMOS) products to obtain more accurate SM data over an area of 2452 km^2 in China at finer spatial resolution (at 1 km-scale) [35]. The authors attempted to predict local SM data—after being downscaled from SMAP/SMOS using RF—from a set of predictor variables at finer resolution, involving vegetation ($NDVI$, ALB, LAI), land surface (LST), hydro-climatic (ET_o), and topographic (DEM, and SL) features. Their approach involved three main steps: (i) resample predictors to coarser resolution of the SMAP and SMOS data and establish a regression relation between the upscaled predictor variables and SMAP/SMOS SM data at coarser resolution; (ii) resample the residuals at the coarse resolution and RF-predicted SM data to finer resolution (1-km scale); and (iii) predict SM at finer resolution from predicted variables at finer resolution using the RF regression developed in step (i) and add the residuals computed in (ii) to the predicted SM data at finer resolution to determine local-scale SM. The authors concluded that RF-downscaled SMAP data performed better than SMOS data. RF-SHAP (a XAI model) analysis unveiled that ET_o, DEM, and ALB were the most influential features for SMAP-RF while ET_o, $NDVI$, and LAI were the most critical features for SMOS-RF. This study introduced a new practical approach for XAI-based local scale SM predictions. It would be beneficial to look into if prediction accuracy of the proposed downscaling method could further improve, if different IAI models other than RF are used.

Remote sensing techniques, however, capture only near-surface SM features and storage [64], which could differ from SM at deeper depths in trends and magnitudes. Therefore, in situ SM measurements were combined with remotely sensed terrain attributes to predict soil-water storage at uninstrumented regions in a basin in California, USA [65]. The authors used the RF model to predict daily inter- and intra-annual SM storage at 10-, 30-, 60-, and 90-cm depths for 6 years, using soil (ST), topographic (TWI, Nt, DEM), and vegetation ($LCON$) features as the predictors. Based on this IAI-modeling set-up, the authors concluded that different predictors were more influential in different periods such as wet-up, snow cover, recession, and dry periods. For example, although ST was consistently a critical feature in all periods, Nt peaked during the wet-up period while DEM and TWI peaked during the recession and dry periods. However, the chosen five predictors were static variables without temporal components, which were used to predict temporal variations in SM at different depths. Inclusion of other time-variant predictors such as P, snow-pack depth, ET in the IAI model could have captured temporal variations in SM predictions more accurately.

Moreover, root zone soil moisture ($RZSM$) is a critical variable for agricultural productivity, crop water stress, and drought monitoring. Accuracy of the optimized RF and physics-based (HYDRUS 1D [66] with data assimilation) models was evaluated for interpolation (for data imputation) and extrapolation (for predictions using testing data) of daily $RZSM$ from a list of predictors, including meteorological (P, T_{min}, T_{max}, U_w, R_s, RH, ET) and vegetation (LAI, CT) features, SM at 5 cm-depth at 15 locations over∼32 month, lagged values of the SM and meteorological variables, in addition to day of the year [67]. The data length was relatively short, yet 50% of the data was allocated to train the RF model. The authors assessed the importance of the variables using the permutation method, which revealed that surface soil moisture, soil properties, and land cover types have larger impacts on $RZSM$ than meteorological variables. Different from earlier IAI-based analyses, the authors compared the performance of the IAI models over the entire period as well as for the extreme dry and wet conditions. They concluded that RF interpolations for $RZSM$ have higher accuracy than RF extrapolations. Moreover, RF interpolations exhibited better prediction accuracy than HYDRUS 1D simulations, but RF extrapolations were comparable to HYDRUS 1D simulations. However, RF overestimated extreme dry conditions, but underestimated extreme wet conditions. This could be due to the relatively short time period used in the analysis, which possibly did not provide enough data to train the model for the extreme conditions properly. Nevertheless, the study demonstrated that the RF model emerged as a computationally efficient prediction tool as an alternative to the Hydrus 1D model to predict $RZSM$.

4.2.4. Groundwater Potential Predictions

Assessment of groundwater potential (GWP) is critical for conservation, sustainable water management, and drought mitigation strategies [68,69]. GWP has been predicted in data-scarce regions using AI models trained by groundwater level, spring inventory, meteorologic, topographic, geologic, soil and surface water data at nearby sites. Predictors used for IAI-based GWP predictions in aquifer data-scarce regions in recent studies are summarized in Table 2.

Information from a limited number of groundwater well locations has been used in recent studies to predict regional-scale GWP in data-scarce regions. The RF and GBoost models were used to predict GWP categorically over a 3339 km^2 region in India using meteorologic (T_a and P), topographic (AL, SA, SD, PlC, PrC), soil (TWI, $NDVI$, ST, $LCLU$), distance ($DisR$, $DisRd$), and geologic (LG) features as the predictors [70]. The IAI models were trained and tested using target data from an equal number of groundwater wells and non-groundwater locations. By allocating 80% of the data for model training, the IAI models produced sensible predictions, where GBoost outperformed RF, and PrC, $DisR$, $NDVI$, and TWI emerged as the most critical features based on the Gini index analysis. In a similar study, the RF, GBoost, and XGBoost model were implemented using geologic, hydrologic, topographic, and land cover features to predict categorically GWP at sites with no wells in an attempt to generate regional GWP maps over an area of 747 km^2 in South Korea [71]. Information from an equal number of groundwater well locations and non-groundwater locations were used to train and test the IAI models, in which 70% of the data was used for training. The authors implemented the elastic net method a priori to eliminate insignificant features to GWP predictions. As a result, they only considered topographic (AL, SD, SA), surface water (DD), soil (TWI), distance ($DistR$, $DisL$, $DisF$), geologic (LG), and soil ($LCLU$, TWI) features as the predictors in the IAI models. Thus, different from [70], the refined predictor list did not include meteorologic variables and ST. The reduced number of predictors used in all three IAI models produced reliable GWP maps for the study area, where XGBoost performed better than GBoost (the second best) and RF. Better performance of XGBoost over GBoost was attributed to (i) implementation of the second-order derivatives in XGBoost—as opposed to first-order derivatives in GBoost—to minimize the loss function and obtain more accurate tree and (ii) regularization features implemented in XGBoost to avoid overfitting.

Table 2. Factors and predictors used in IAI-based groundwater potential predictions.

Factors	Predictors
Meteorologic	Temperature (T_a), Precipitation (P)
Topographic	Altitude (AL), Slope aspect (SA), Slope degree (SD), Slope length (SL) Convergence index (CI), Plan curvature (PlC), Profile curvature (PrC), Relative slope positioning (RSP), Vertical distance to channel (VDC), Terrain ruggedness index (TRI), Melton ruggedness number (MRN), Multi-resolution ridge top flatness (MRRTF), Multi-resolution valley bottom flatness (MRVBF)
Geologic	Lithology/geology (LG), Fault density (FD), Lineaments density (LD)
Surface water	Drainage/river density (DD)
Soil moisture, Surface	Soil texture (ST), Stream power index (SPI), Topographic wetness index (TWI), Normalized difference vegetation index ($NDVI$), Land cover/use ($LCLU$)
Distance from man-made or geologic structures	Distance from river/drainage ($DisR$), Distance from lineament ($DisL$), Distance from road ($DisRd$), Distance from fault ($DisF$)

In the absence of detailed aquifer and groundwater level data, spring data have been used as a surrogate predictand to estimate GWP. The optimized parallel RF (PRF) and XGBoost were used to determine GWP categorically in data-scarce regions in Iran on the basis of spring data using only DEM-derived spring associated factors (DEM-SDF) [72]. These factors included topographic (AL, SA, SD, SL, CI, PlC, PrC, RSP, VDC), surface water (DD), soil/surface (SPI, TWI), and distance ($DistR$) features, which were used as predictors. The authors used 944 springs locations and randomly generated 944 non-spring locations over an area of 1676 km^2 as the target data. Based on the 70:30 split ratio for the training and testing datasets, the authors reported that PRF and XGBoost predictions showed∼80% similarity and predicted high GWP regions closely. Different from the conclusion in [70], Gini impurity revealed that CI, TWI, RD, and AL are the most indispensable features in GWP predictions based on spring data and DEM-SDF.

Similarly, an AI-driven regional GWP map was developed based on the spring data [73]. The authors used the optimized RF, LR, Decision Trees (DT), Artificial Neural Networks (ANN), and their combinations (i.e., additional 11 AI models) to predict GWP categorically over a karstic aquifer in a mountainous region in Morocco using the spring inventory as the target variable, and meteorologic (P), topographic (AL, SA, SD, SL, CI, PlC, PrC, TWI, TRI, MRN, $MRRTF$, $MRVBF$), soil/surface (SPI), geologic (LG, LD, FD), distance ($DisF$, $DisL$, $DisR$), surface water (DD), surface and soil moisture-related ($NDVI$, $LCLU$) features as the predictors. The spring inventory data included 347 spring locations and 1124 randomly chosen non-spring locations. They allocated 75% of the data to train the models. Prior to AI analysis, they performed multicollinearity analysis to determine linear dependency among the predictors to avoid redundancy, and computed information gain (IG) to identify the predictors positively associated with the enhanced GWP to reduce the number of predictors. However, multicollinearity analysis is not required for IAI modeling, as the models can handle redundant predictors. Besides, RF-SHAP (a XAI model) can unveil more effectively and accurately the order of importance of the predictors and the inflection point of each predictor above/below which the predictor would result in enhanced or reduced GWP. Nonetheless, based on multicollinearity and IG analyses, the authors retained all the predictors in the AI analysis. Based on RF-driven ranking, LG (lithologic), FD, DF (tectonic), P (meteorologic) were identified to be the most important predictors. Different from the analysis in [72], the authors tested the predictive accuracy of the weighted-aggregation of RF, LR, DT, and ANN to estimate GWP, where the weights were set to the area under the success rate curve from each AI model. They concluded that weighted-average RF-DT and RF-LR-DT (IAI models) yielded the best prediction accuracy for GWP prediction for the semi-arid karstic mountainous region.

The results from the studies discussed above are based on different sets of mostly static region-specific predictors. Therefore, it is difficult to make generalizations over relative predictive accuracy of the IAI models used. The IAI-based predictions discussed above require a priori domain knowledge of the variables and system, as the AI predictions based on DEM-SDF would be applicable only to basins where GWP is expected to be controlled largely by topographic features. Conversely, the topographic watersheds of karst catchments have little significance for their aquifers [74], therefore such IAI models may not be applicable to estimate GWP in karstic aquifers. The predictors in the studies discussed above did not include geospatial information about aquifer characteristics such as aquifer type, aquifer thickness, depth to water table, and aquifer parameters (e.g., transmissivity or storativity) in predicting regional GWP due to scarcity of data, although these features strongly determine GWP and productivity of aquifers. Furthermore, although well-balanced datasets were used in [70–72], an imbalanced dataset (1:3 ratio) was used in [73]. Imbalanced datasets in model training, however, could cause bias towards the minority class, and hence, impair the prediction accuracy of the model.

4.2.5. Groundwater Level Predictions

Groundwater levels (GWL) could be affected by climate factors, land use, pumping, and hydraulic interaction with surface and other subsurface waters. Short-term GWL predictions could be imperative for landslide prone areas [75], in agricultural regions for scheduling irrigation [76] and in regions that experience sudden increase in groundwater withdrawals or extreme climate events (e.g., heatwaves). Long-term GWL predictions under future climate scenarios are critical for development of sustainable groundwater management plans [18] and sustainability of agricultural production systems [77]. Predictors used for IAI-based GWL predictions in recent studies are summarized in Table 3.

Table 3. Factors and predictors used in IAI-based groundwater level predictions.

Factors	Predictors
Meteorologic	Precipitation (P), Temperature (T_a), Solar radiation (R_s)
Hydrologic	Lagged GWL, Terrestrial water storage (TWS)

Using meteorological data, the optimized XGBoost, RF, and SVM models, and their hybrid versions were implemented with or without wavelet transforms (WT) for short-term monthly GWL (1–3 months ahead) predictions in Kumamoto City, one of the regions with the highest groundwater use in Japan [78]. The authors used monthly time-lagged GWL, monthly-average T_a, average monthly total P, and cumulative monthly P as the predictors. WT was used to extract time-variant information such as trends and periodicity in the AI modeling. However, such time-variant domain knowledge can alternatively be incorporated using day, month, or year as the engineered features in the AI modeling. The authors used 442 records, and implemented a 85:15 ratio for the training and testing datasets. They concluded that SVM outperformed XGBoost and RF, when the WT is not included. When the AI models were coupled with the WT, however, SVM and XGBoost exhibited comparable predictive accuracy while outperforming RF. WT-AI coupling apparently enhanced the prediction by 3–5%, which was more beneficial for 2–3 months ahead predictions. The authors adopted minimal-redundancy-maximal-relevance to rank the order of importance of the predictors.

The effectiveness of an optimized hybrid K-Nearest Neighbors (KNN)-RF model (a coupled non-IAI and IAI models) for short-term prediction (2 weeks to 3 months ahead) of daily GWL in a near-surface aquifer in a data-scarce region in Rwanda [79] was analyzed. The authors used GWL measurements from a single borehole after removing anomalies via a time-series filtering method prior to their use in the AI analysis. GWL was related to T, P, R_s, and their 1–4 days lagged values. The AI analyses were performed using $\sim$2 years of daily data with 759 records. The authors implemented a 'walk-forward' approach to predict GWL from the input data while implementing 88:12 ratio for the initial split for the training and testing datasets. KNN-RF consistently exhibited better prediction accuracy at 15, 30, 60, and 90 days predictions than RF, KNN, SVM, and ANN. Using the KNN-RF model with different combinations of the predictors, the authors concluded that R_s, T, and GWL time-lags in addition to the first lag P were the most influential predictors on short-term GWL forecasts. This could have been more effectively analyzed using RF-SHAP (a XAI model), instead of multiple KNN-RF model runs with different combinations of the predictors. These IAI, non-IAI, and hybrid IAI and non-IAI modeling studies sought to predict short-term GWL based on local meteorologic and hydrologic data. Inclusion of groundwater withdrawals, aquifer parameters, and aquifer recharge in AI-based GWL forecast analysis could increase their wider acceptance by the water resources and hydrology community.

Different from the applications above, the XGBoost, multivariate LR, RF, multilayer perceptron neural network (MLP), and SVR were used for image (map)-based prediction of monthly GWL in the southern regions of the African continent at the pixel-level from monthly terrestrial water storage (TWS) maps, the coordinates of the pixels on TWS maps, and monthly time-stamp [80]. After imputing 10% of the missing monthly images, the authors generated 161 sequences of 12 consecutive images for the period of 2002 and 2019, in which the first 149 images were used for model training and the rest for model testing. The sample size to train the AI models was low in this application. Nonetheless, XGBoost with the gain matrix determined that TWS pixel information from 12-, 11-, and 1 preceding months were the most influential predictors to estimate GWL in the current month. Among the AI models used, SVR reportedly yielded the best prediction accuracy in predicting GWL. In this application, XGBoost (an IAI model) provided the information on the feature importance and selection, and SVR (a non-IAI model) yielded overall better prediction accuracy, similar to the implementation of hybrid IAI and non-IAI models in [34,81]. The use of additional information on spatiotemporal variations in groundwater withdrawals, however, could have improved the accuracy of GWL predictions.

4.2.6. Streamflow Predictions

Streamflow (Q_s) is impacted by climate change and human activities, such as dam construction, changing environment, and increased surface water diversions to meet the consumptive water demands in areas with increasing populations [82]. Predictors used for IAI-based Q_s predictions in recent studies are summarized in Table 4.

Table 4. Factors and predictors used in IAI-based streamflow predictions.

Factors	Predictors
Meteorologic	Precipitation (P), Temperature (T_a)
Hydrologic	Lagged Q_s
Hydro-climatic and soil-associated	Pan evaporation (E_p), Evapotranspiration (ET)
Surface Vegetation	Effective vegetation index (EVI)

Changing climate and intensified human activities could make the relation between Q_s and predictors non-stationary, which was referred to as the concept drift in [83]. Because new climate change and human impacts on Q_s are not captured in historical data used for model training, the AI model would not be informed about such gradual or abrupt unprecedented changes that would violate the stationarity assumption, unless the AI model is 'intervened' and informed of them. The performance of XGBoost with concept drift detection (CDD) was compared against XGBoost without CDD, RF, SVM, and DTR in predicting one-month ahead Q_s at the Qingliu river catchment in China using meteorologic (P, T_a), hydroclimatic (E_p), soil (EVI), and hydrologic (past Q_s) features [83]. In this study, CDD operates based on presumably normally-distributed historical error rate. In XGBoost-CDD modeling, when unprecedented Q_s rates were detected, XGBoost was re-trained with the existing data; otherwise, it was incrementally trained. Using monthly data from 1989 to 2010 and assigning 70% of the initial data for model training, XGBoost-CDD outperformed the prediction accuracy of XGBoost, RF, SVM, and DT, as XGBoost-CDD detected the abrupt change in Q_s in 2003 due to the rapid development of society and economy, quick population growth, and dramatic changes in land cover and use in the region, which required for XGBoost-CDD to be re-trained. In our opinion, the IAI-framework in [83] sets the stage for interventional IAI in hydrologic applications, as hydrological settings would likely expose to unprecedented consequences of human activity and changing climate on Q_s more often in the future.

The optimized XGBoost-Extreme Learning Machine (ELM) model was used to predict one-month ahead monthly Q_s in the Göksu-Himmeti catchment area in Turkey from hydrologic (multi-lagged Q_s), meteorologic (P, T_a), and hydroclimatic (ET) data from 1973 to 2010, in which 75% of the data was used for model training. In this study, XGBoost was used as the feature selection tool and ELM as the predictor tool. The authors used 'gain score' in splitting a leaf into two leaves in XGBoost to determine the most influential lags among 30 lags for each predictor. After testing XGBoost with different combinations of multi-lagged predictors, Q_s, P, and T_a were reported to be the most critical features for one-month ahead Q_s prediction for the study area. The feature importance ranking was used to select the features for ELM. The authors concluded that XGBoost-ELM (a hybrid IAI and non-IAI model) provided higher predictive accuracy than XGBoost alone. Similar to [34], the advantages of the IAI and non-IAI models were combined in [81] to achieve higher predictive precision of Q_s.

In addition to IAI-based predictions of Q_s from meteorologic, lagged hydrologic, hydro-climatic, soil-associated, and land surface features, IAI-based models were used to predict Q_s from its spectral and frequency components. For example, singular spectrum analysis (SSA) and LGBoost were integrated to predict real-time urban runoff in Yuelai New City in China [84]. The authors used 39 rainfall events in this study, in which 33 of them were used for model training and 6 of them for testing. After extracting the trend, fluctuation, and noise components from the runoff time series using SSA, they reconstructed the series using LGBoost. The motivation was that the data pre-processed with the SSA, or other decomposition methods, could significantly improve the AI performance. The authors noted that SSA-LGBoost predicted runoff with higher accuracy and peak error <18%, outperforming LGBoost and LSTM models. On the other hand, using the Fourier Transform (FT) to decompose 10-day inflow time series, the performance of XGBoost and SVR was tested to forecast the decomposed components, based on frequency domain analysis, with each component comprising contiguous frequencies and exhibiting a clear physical meaning [85]. The authors used the Three Gorges Dam inflow series in China. The 10-day records from 1990 to 2009 were used for training and 2010 to 2015 for testing. Three decomposition strategies were tested: The centered 10-day inflow time series (only one decomposed component) and decomposition into four and seven components. Their results showed that FT-SRV almost perfectly derived the 10-day streamflow forecast with 7 components and outperformed the other decomposition approaches. In addition, their analysis showed that the FT-XGBoost presented a worse performance than the FT-SRV.

4.2.7. Water Level Predictions in Reservoirs, Lakes, and Delineation of Wetlands

Lakes and reservoirs are important fresh water sources for domestic, industrial, agricultural, and recreational water uses, regional flood control, and aquaculture [86]. Water level (WL) is an important physical indicator of lakes, and its fluctuations may impact the sustainability of lake ecosystems [87], and consumptive water uses under current and future climate conditions and human activities [88]. Similarly, wetlands are a critical component of a hydrologic system for maintaining hydroecology, flood control, providing nutrients, and controlling WL in surface water systems [89]. Predictors used for IAI-based predictions of WL in reservoirs, lakes, and for delineation of wetlands in recent studies are summarized in Table 5.

Table 5. Factors and predictors used in IAI-based water level predictions.

Factors	Predictors
Meteorologic	Precipitation (P), Temperature (T_a), Wind speed (U_w) Standardized precipitation index (SPI)
Topographic	Digital elevation model (DEM), Slope (SL)
Geologic	Lithology/geology (LG)
Hydrologic	Lagged WL, Downstream releases (Q_{DR}), Water table (WT), Aquifer permeability (K_s)
Soil, Surface	Soil texture/type (ST), Topographic wetness index (TWI), Normalized difference vegetation index ($NDVI$), Flow accumulation factor (FAP)
Site Specific	Water levels at the embankment, Drainage pump station, Surface water abstraction (Q_{SW})

In regards to WL predictions in reservoirs, the optimized Boosting, RF, Bayesian Linear (BL) and Neural Network (NN) model were used to predict a day- or week-ahead WL in the Keymir reservoir in Malaysia, operated for hydropower generation [90]. Two scenarios with a small number of predictors were considered. In the first scenario, daily P and WL from 1985 to 2019 were used as the predictors to estimate WT. In the second scenario, daily Q_{DR} from 2010 to 2019 was also used as the predictor. Using 80% of historical data to train the AI models, the authors achieved higher prediction precision for a day- or week-ahead reservoir WL when they included Q_{DR}, where the prediction accuracy of the AI models ranked in the order of Boosting > RF > BL > NN. In this study, IAI models performed better than non-IAI models in predicting short-term WL in a reservoir. The authors performed sensitivity analysis to assess prediction uncertainties. This could have been alternatively achieved by combining the Boosting model with the NGBoost as in [17]. If additional predictors (e.g., ET, more lags in WL, P) are used in such analysis, SHAP analysis can be used to identify the most influential predictors to reduce the input dataset for the IAI and XAI modeling.

Aside from WL predictions in reservoirs and lakes, the optimized RF model was used to infer the importance of climatic and abstraction features on WL fluctuations in Lake Bracciano in central Italy, which is designated as an emergency water source to be used in severe droughts [91]. The authors resorted to the IAI modeling, as they did not have sufficient data on water exchange rates between groundwater and lake to construct a lake water-balance equation or use physics-based models. They analyzed the influence of short-term (e.g., run-off) and long-term (e.g., groundwater dynamics) effect of the monthly P using SPI at different time scales (1–24 months), ET_o, T_a, Q_{SW}, and month of the year on WL for the period of 1955–2019. Using 50% of the data for model training and implementing computationally-expensive drop-column feature importance approach, they concluded that $SPI24$, Q_{SW} month of the year, $SPI12$, $SPI13$, $SPI6$, and T_a were the most critical features. This suggests that P associated with the groundwater dynamics and water abstraction were the most influential process while T_a was the least critical variable. Using the RF model, the importance of Q_{SW} with respect to long-term P variability was shown to increase by 15% after 1985. The authors noted that the importance of a month index needs to be analyzed in combination with the associated time scale of the P anomaly. This and the feature importance analysis can be done effectively using the local and global SHAP analysis. The SHAP analysis can also reveal the effect of percentage increases or decreases in the predictors' values (e.g., $SPI24$ and Q_{SW}) on WL fluctuations, which are imperative to assess the potential impacts of changing climate and water abstraction policies on WL.

As for WL predictions in wetlands, RF, DT, SVM, and ANN were used to predict daily WL in Upo wetland in South Korea, which is a large inland wetland with high biodiversity [92]. The predictions were based on 1–3 days lags of minimum, maximum, and average T_a, P, maximum and minimum U_w, and WL at the nearby embankment and drainage pump station. Using the measurements from 2009 to 2015 and keeping the data from the last two years for the model testing, the authors concluded that RF outperformed DT, SWM, and ANN in predicting the overall trend, peak values, and peak occurrence times of WL. They also noted the need for further improvements in peak value predictions and peak delay error reductions, which could be achieved by accommodating the information on soil characteristics, GWL, and backflow during rainy seasons if/when such data are available. Through the degree of increases in the node purity in the RF-based modeling, the authors identified 1–3 days lags in WL at the nearby embankment, 1-day lag in P and in WL at the drainage pump stations were the most critical features in predicting WL at the wetland. Alternatively, RF-SHAP (a XAI model) could have been used for the feature importance ranking.

In addition to the use of individual AI models, multilayer pattern recognition tools based on multiple supervised AI models have been developed to construct predictive maps based on point-source observations. As such, MLMapper is a AI-based predictive map development tool that performs predictive analyses using 20 different AI models (including IAI and non-IAI models) and site-specific predictors. MLMapper was used to delineate the surface area of groundwater-dependent ecologically-sensitive wetland areas in central Spain, using information on geologic (LG), hydrologic (WT, K_s), topographic (DEM, SL), and soil ($NDVI$, FAP, TWI, ST) features [93]. The data size, however, was low for a typical AI modeling, which consisted of 75 known wetland points and 75 non-wetland points. The authors varied the split ratio for the training and testing data from 50:50 to 80:20. They concluded that tree-based models (ERT, RF) outperformed most other supervised classifiers in terms of raw test score, surface area, and number of explanatory variables required for mapping. Trained AI models predicted larger wetland surface areas than the natural inventory, suggesting that a combination of the features identified additional wetland areas not captured in field surveys. Although MLMapper reportedly performs a collinearity test to identify and eliminate redundant features, this is not a requirement for tree-based IAI models. Weighing and permutation importance methods used with the ERT and RF revealed that DEM, LG, WT, and K_s were the most influential features in determining the spatial extent of wetlands. However, ERT-SHAP or RF-SHAP (XAI models) could have been used instead to rank the most influential features without resorting to the recursive feature elimination methods implemented with MLMapper. Moreover, local SHAP analysis could have been used to determine the inflection points above or below which the predicted wetland surface area (represented as a binary variable) may increase or decrease with changes in predictors' values. Therefore, we expect to see the use of the local and global SHAP analyses in such automated AI-based predictive map construction tools in the near future.

4.2.8. Water Quality Predictions

Prediction of salinity and pollution levels of surface water and groundwater, and identification of the most critical physicochemical parameters affecting local and regional water quality are imperative for their sustainable operations and well-being of aquatic ecology [94]. Predictors used for IAI-based water quality predictions in previous studies are summarized in Table 6.

Table 6. Factors and predictors used in IAI-based water quality predictions.

Factors	**Predictors**
Physicochemical	pH, Total dissolved solids (TDS), Total hardness (TH), Cations, Anions, Total phosphorus (TP), Nitrate concentration (C_{NO_3}), Nitrite concentration(C_{NO_2}), Pesticide concentration (C_P), Biochemical oxygen demand (BOD), Chemical oxygen demand (COD), Electrical conductivity (EC), Nitrate-Nitrogen ($NO_3 - N$), Nitrite-Nitrogen ($NO_2 - N$), Phosphate (PO_4^{3-}), Surface water temperature (T_{sw}), Turbidity (NTU), Dissolved oxygen (DO)
Meteorologic	Precipitation (P), Temperature (T_a)
Hydro-climatic	Evaporation (ET)
Hydrogeologic	Aquifer type, Aquifer transmissivity (Tr), Horizontal hydraulic conductivity (K_h), Vertical hydraulic conductivity (K_v), Aquifer thickness (A_t), Aquitard thickness (Aq_t), Depth to water table (DWT), Groundwater level (GWL), Depth to screen well (DSW), Pumping capacity ($Q_{p,max}$), Well density (WD), Pumping density ($Q_{P,d}$), Operation time-length of wells (OW), Aquifer recharge (AR), Soil type (ST)
Hydrologic	Streamflow (Q_s) , Stream length (ST_L)
Topographic	Altitude/elevation (AL), Land surface slope (SL)
Land use	Crop type, forest, urban residential land, pasture land
Site and problem-specific	Presence of streams, distance from sea ($DisS$), Population density (PD), Distance to saline sources ($DisSS$), Distance to fault (DF)

The water quality index (WQI), which integrates several physical and chemical factors into a single parameter, has been commonly used to evaluate or categorize the quality of groundwater and surface waters [95]. The predictive performance of the optimized RF, XGBoost, ANN and DL models was analyzed in determining entropy weight-based groundwater quality index (EWQI) in the Mahanadi basin in India from a set of physicochemical parameters, involving pH, TDS, TH, Ca^{2+}, Mg^{2+}, Na^+, K^+, HCO_3^-, Cl^-, SO_4^{2-}, NO_3^-, F^-, and PO_4^{3-} [96]. The authors applied the AI models with the data from 226 locations. They varied the split ratio for the training and test data from 75:25 to 85:15 to seek the best prediction accuracy. Although the authors noted that data normalization should be performed prior to these AI models, data normalization is not required for RF and XGBoost. The predictive performance of the AI models were reported to be in the order of DL > XGBoost > ANN > RF, in which DL (a non-IAI model) yielded better predictive accuracy than XGBoost (an IAI model), yet it was unable to unveil the reasoning behind the predictions. Therefore, the authors resorted to inter-criteria correlation to determine the order of importance of the predictors. However, this could have been accomplished by XGBoost-SHAP (a XAI model), which can also provide inflection point values for each predictor above or below which EWQI would increase or decrease.

Groundwater salinity is a critical water quality measure that could affect sustainable use of inland or coastal aquifers. The optimized XGBoost, multiple linear regression (MLR), and DNN models were used to map spatial distribution of groundwater salinity, described in terms of EC, in a coastal aquifer of the Caspian Sea in Iran using data from 140 piezometric wells [97]. The authors used a 75:25 split ratio for the training and test data. Hydrogeologic (Tr, DWT), site-specific ($DisS$), meteorologic (mean annual P), hydro-climatic (ET), topographic (AL, SL) features were initially considered as the predictors. The authors used the MLR model to identify the contribution of each predictor to EC in a stepwise manner by adding and removing each predictor to MLR until they reached the maximum predictive accuracy on the test data. Based on the MLR analysis, ET and SL were removed from the predictors list due to their negligible contributions to EC. However, use of a linear model to rank the importance of the predictors and remove the least important ones from the AI models for a nonlinear problem is questionable. XGBoost-SHAP (a XAI model) would have been an accurate and robust choice to rank the importance of the predictors for such nonlinear problems. Nonetheless, their analysis unveiled that the predictive performance of the AI models was in the order of XGBoost>DNN>MLR on the test data, indicating that XGBoost (an IAI model) exhibited higher prediction accuracy than DNN (a non-IAI model) in groundwater salinity predictions in a coastal aquifer.

The optimized CatBoost, XGBoost, LGBoost, and RF models were used to predict groundwater salinity in a multilayer coastal aquifer over an area of 3312 km^2 in the Mekong Delta, Vietnam [98]. Using 216 groundwater samples taken in rainy and dry seasons from 2013 and 2018 with the influencing factors, including site-specific ($DisSS$, DF, $DisS$) and hydrogeologic (DSW, GWL, K_h, K_v, Aq_t, $Q_{p,max}$, WD, OW, $Q_{P,d}$, ST) features, and assigning 70% of the data for model training, they concluded that the predictive accuracy of the AI models was in the order of CatBoost > XGBoost > RF > LGBoost. Importance of the predictors was determined using the CatBoost ranking. As a result, ST, $Q_{P,d}$, and $DisS$ were removed from the predictors list. The reduced input set enhanced the prediction accuracy of CatBoost slightly. Although the authors normalized the predictors prior to AI analysis, such normalization is not required for these IAI models. Using the CatBoost model, the authors constructed a regional groundwater salinity map based on the predicted chloride concentrations, which unveiled that paleo-saline groundwater salinization is the main process for increased salinity in the study area and identified salinity-affected populations. Thus, the IAI modeling in this study raveled information not only about salinity intrusion mechanism, but also on its social dimension.

Vulnerability maps have been used to identify areas most vulnerable to water quality deterioration. Index-based techniques have been widely used for preparation of groundwater vulnerability assessments maps due to their computational simplicity and less data demand compared to statistical or process-based simulation techniques [99]. GALDIT is an index-based method to assess groundwater vulnerability to saltwater intrusion using information on hydrogeologic (aquifer type, A_t K_h, DWT), site-specific ($DisSS$) features, and impact of existing seawater intrusion status. The main drawback of such index-based methods is the subjectivity of each variable's rating and weight in estimating the vulnerability index. The optimized XGBoost, LGBoost, Adaptive Boosting of Decision Trees (AdaBoost), CatBoost, and RF were used to overcome the subjectivity of the weights and ratings assigned to the variables in the GALDIT framework when calculating a groundwater vulnerability index over 500 km^2 area in the Lake Urmia catchment area in Iran [100]. The authors implemented boosting aggregation (bagging) and disjoint aggregation (dagging) sampling methods to increase the data size and reduce the prediction variance in this study area with the initially small data size. The GALDIT indices, after being adjusted using TDS measurements, were used as the predictand. Using 70% of the data to train the AI models, the authors concluded that the precision accuracy of the models was in the order of XGBoost > AdaBoost > RF > CatBoost > LGBoost when a bagging or dagging resampling method was not implemented. Although these IAI models improved the prediction accuracy of groundwater vulnerability by ∼15% in comparison to standard GALDIT framework, and additional precision enhancement of ∼5% was achieved using bagging-XGBoost, the final prediction accuracy was however not statistically significant. The authors noted that the six predictors implemented in the GALDIT framework may not be sufficient to determine groundwater vulnerability. They also noted that the AI models chosen in their study cannot suggest a new weight or rating score for each variable because the IAI models are 'black box' models. This statement is, however, questionable. Although ensembling can make AI model interpretability and explainability harder, the boosting and bagging AI models are not black-box models, as they can readily be coupled with the explanatory methods for enhanced interpretability (IAI models) and explanability (XAI models), as discussed in Section 2. In fact, information on new weights and ratings to enhance the groundwater vulnerability index can be obtained by coupling the bagging and boosting AI models with the SHAP method, as in [18].

Chemicals used on farmlands pose risks on the water quality and human health, and environmental and ecological well-being. The optimized XGBoost, ANN, and SVM were implemented to predict nitrate and pesticide concentrations in groundwater (regression problem) and associated risk (classification problem) using hydrogeologic features (e.g., aquifer type and properties), land use features (e.g., nearby croplands, forest areas, primary water uses), and water quality measures (C_{NO_3}, C_P), and other physicochemical parameters [33]. The analysis was conducted in a data-scarce region, involving 303 sampling wells from 12 midcontinental states in the USA, and 80% of the data was used for model training. The authors analyzed imbalanced classes using 'confusion matrices' in classification problems and implemented oversampling and cost-sensitive learning to address the problem of imbalanced classes. They noted that ANN performed better than XGBoost for the regression task, but XGBoost produced a majority of the best predictions among all three models. In addition, unlike the ANN and SVM models, XGBoost-SHAP (a XAI model) identified the order of importance of the predictors influencing the nitrate and pesticide concentrations and associated risk classifications and concluded that both nitrate and pesticide were the most important predictors of each other. In another study, the RF and MLR models were used to explain groundwater C_{NO_3} contamination at the African continent-scale in relation to land use, soil type, hydrogeology (aquifer type, K, DWT, AR), topography, climatology (climate and rainfall class), nitrogen fertilizer application rate, and PD in the absence of a systematic groundwater monitoring program [101]. The analysis focused on spatially-variant mean C_{NO_3} without addressing their temporal variability. Using 80% of the data for model training, the authors concluded that RF outperformed MLR in predicting C_{NO_3}. The main advantage of RF (an IAI model) over MLR (a statistical model) is that RF is (i) a non-parametric model, i.e., the model structure does not need to be specified a priori, (ii) more efficient in determining nonlinear relationships and patterns between target and multidimensional predictors without relying on restrictive assumptions such as particular statistical distribution for residuals, non-collinearity among the predictors, (iii) as interpretable as MLR yet provide better predictive accuracy, and (iv) a robust model for outliers. These advantages are equally applicable to other tree-based ensemble AI models.

As mentioned above, WQI has also been used to assess water quality in stream waters. The ERT, DT, and SVM models were used to predict WQI at the Lam Tsuen River in Hong Kong from a set of physicochemical features [102]. Monthly physicochemical features included pH, BOD, COD, DO, EC, $NO_3 - N$, $NO_2 - N$, PO_4^{3-}, T_{sw}, NTU from 1998 to 2017. The author noted that when these 10 features were used as the predictors, the prediction performance of the AI models was in the order of ERT > SVM > DT. By trying different combination of the predictors, ERT (an IAI model) with the reduced list of predictors, including only BOD, PO_4^{3-}, and NTU achieved the second best prediction accuracy. Thus, in the absence of a full set of physicochemical data, ERT with BOD, PO_4^{3-}, and NTU could still provide a good estimate for WQI for surface waters. However, instead of manually trying different (and gradually reduced) combinations of predictors in search of high WQI precision, ERT-SHAP (a XIA model) can be used to identify high fidelity ERT models with the least number of predictors.

Hybrid physics-based and AI models have also been used to predict water quality measures. The XGBoost model was combined with the Soil and Water Assessment Tool (SWAT) [103] to estimate TDS and better understand water salinity river in a semi-arid agricultural Rio Grande Watershed in Texas [104]. XGBoost was trained with water quantity and quality data that were monitored in nine locations. The predictors used in their study were physicochemical (C_{NO_2}, C_N, TP), meteorologic (P), topographic (AL), and hydrologic (Q_s, ST_L, dominant ST). Results from calibrated the SWAT model were used as inputs to XGBoost to predict TDS. However, the SWAT model could not be properly calibrated for all studied locations due to a lack of data. In addition, the insufficient data compromised XGBoost training and caused overfitting. These conclusions highlight the importance of high-quality and sufficient data for proper analysis with AI. The authors also argued that if additional water quality parameters are monitored, more predictors could be used and the results would be more accurate. Despite the insufficient data, the AI modeling approach showed to be advantageous over simple SWAT modeling as it improved the bias and variance of TDS estimates.

In addition to physicochemical parameters, water surface temperature (T_{sw}) is an influential factor for water ecosystems and, hence, for successful water management plans. The performance of five AI models were compared to predict T_{sw} of 25 lakes in Poland [105]. The analyzed models were ERT, multivariate adaptive regression splines (MARS), M5 Model tree (M5Tree), RF, and MLP. Although AI models have been successfully used in broad hydroclimatic applications, none of the AI models in [105] were able to outperform prediction accuracy of the physics-based 'air2stream' model [106]. The authors suggested including more predictors to potentially improve the prediction accuracy of the AI Models.

As for potential future directions, IAI and XAI can be used to examine how DO, T_{sw}, total and reactive iron (Fe), redox potential, and sulfate (SO_4^{-2}) and associated biogeochemical processes [107,108] in freshwater environments could vary with the depth in response to changing hydroclimatic conditions under future climates. This could be useful to predict the depths at which aerobic and anaerobic processes prevail, which would have direct impacts on future aquatic ecology and consumptive water use. In addition, infiltration of micro and nanoplastics into freshwater environments is becoming a growing concern worldwide [109,110]. When more regional and global data become available, IAI and XAI models could be useful to analyze the relative importance, interdependency, and interaction of environmental factors (e.g., minerals, pH, natural and dissolved organic matter, ionic strength, net surface charge of plastics [111]) on the the fate and transport of micro and nonoplastics in aquatic environments and consequently their ecological impacts under different hydroclimatic conditions.

4.2.9. Flood Hazard Risks Prediction

Floods are caused by heavy rainfall over lowlands with gentle slope and low water infiltration capacities that can be accompanied by debris flow and landslides. Floods often cause many casualties and property losses. Such extreme events are expected to occur at higher frequencies in a globally warming climate and due to intensified human activities [112]. Flood risk assessments are important for flood insurance, floodplain management, and disaster warning systems. AI-based flood predictions and risk assessments so far typically focus on passive predictions without considering adaptation measures and resilience of social and economic dimensions. The predictors used in recent IAI-based flood forecast analysis are summarized in Table 7.

Hydrodynamic models are commonly used for the flood managements. These models solve complex physical equations to estimate floodplains, which makes them computationally inefficient, especially two-dimensional (2D) models. This drawback prevents the application of such models to a large-scale domain, and AI can be an alternative. For example, the RF and MLP models were combined for fast water depths predictions [113]. RF was applied to identify wet (flooded) and dry cells using flow and the domain coordinates as inputs. Then, MLP used RF's output to compute river depths in the wet nodes. The authors used the International River Interface Cooperative software (iRIC) model with FaSTMECH (Flow and Sediment Transport with Morphological Evolution of Channel) solver [114] for hydrodynamic modeling, which was calibrated and used to train the AI models. Seven events with different flow magnitudes (10, 50, 95, 120,150, 300, and 400 m^3/s) were used for training and five events with different flow magnitudes (20, 30, 45, 225, and 350 m^3/s) were used for testing. This approach was evaluated in Green River in Utah, USA and was able to reduce the simulation time by 60 times with satisfactory prediction performance. However, the method was tested for a single location, and its prediction capabilities to other reaches still need to be evaluated. Generalization to different areas is essential for the applicability of such models to large-scale domains.

Table 7. Factors and predictors used in IAI-based flood hazard risk predictions.

Factors	Predictors
Disaster-inducing factors	Maximum 3 day precipitation ($M3DP$), Maximum 3 h precipitation ($M3HP$), Maximum 1 day precipitation ($M1DP$), Annual P, Days with precipitation exceeding 25 mm ($DPE25$), Precipitation of the wettest month (P_{wm}), Precipitation of the driest month (P_{dm}), Precipitation seasonality (P_s), Precipitation of the wettest quarter (P_{wetq}), Precipitation of the driest quarter (P_{dryq}), Precipitation of the warmest quarter (P_{wq}), Precipitation of the coldest quarter (P_{cq}), Percentage of the catchment area affected by rain (PAA), Typhoon frequency (TF), Streamflow (Q_s), Runoff depth (RD)
Disaster-breeding environmental factors	Slope (SL), Digital elevation map (DEM), Altitude (AL), Distance to river ($DisR$), Land use patterns (LUP), Normalized difference vegetation index ($NDVI$), Road density (RD), Soil texture (ST), Soil depth ($SDep$), Soil moisture (SM), Topographic wetness index (TWI), Curve number (CN), Stream power index (SPI), Vegetation coverage (VC), Lithology/geology (LG), Distance from road ($DisRd$), Profile curvature (PrC), Plan curvature (PlC), Hillshade (HS), Flow accumulation (Q_{ACC}), Slope aspect (SA), Vertical flow distance (VFD)
Disaster-bearing body factors	Population (POP), Population density (PD), Gross domestic product (GDP), Gross domestic product density ($GDPD$)

The performance of RF to predict runoff discharge was compared against the 'hydromad' hydrological model [115] for 95 basins in the USA and Canada [116]. In this study, P, T_{max}, T_{min} and SM were used as the predictors. In addition, the effects of catchment characteristics were also evaluated by including additional predictor variables, such as the standard deviations of P, T_{max} and T_{min} within the catchments and PAA. Their results showed that climate conditions and elevation could affect the RF performance. Although the authors noted that RF can be an alternative to traditional hydrological models, they highlighted that RF failed to predict high magnitude flows. In addition, RF only provided robust results for catchments with a warmer climate and lower altitudes. Further research was suggested to increase its accuracy for larger magnitude events and to improve RF prediction capabilities in more heterogeneous catchments. In colder catchments, for instance, the authors suggested including snow and soil moisture as predictors. In semi-arid regions, lack of flood training data compromised the model performance. Their results shows this type of AI is suitable for use in large-scale basins and can improve flood risk assessments at a national or continental scale.

In some other applications, data-driven AI models were used as a sole predictor for flood risk. Current and future flood risk in the Kalvan watershed in Iran was evaluated with AI [117]. The future conditions were evaluated for 2050, with the projected changes in climate and land use. The authors used conditional inference random forest (CIRF), GBoost, and XGBoost to model the flood risk. In addition, a combined prediction with these three approaches was also evaluated. Twenty predictors were used to build the models, including those associated with the disaster-inducing factors (annual P, P_{wetm}, P_{drym}, P_s, P_{wetq}, P_{dryq}, P_{wq}, P_{cq}) and disaster-breeding environmental factors (SL, DEM, AL, DTR, ST, LG, DRd, LUP, PrC, PlC, SPI, TWI). The results indicated that the combined approach had the highest accuracy, followed by GBM, XGBoost, and CIRF. In general, all models attained a satisfactory performance and are suitable for flood risk mapping. Similarly, LGBoost and CatBoost were used to determine flash flood susceptibility and compared their performance with RF [118]. The authors used over 400 flood maps to train and test the models, split in 70% for training and 30% to testing. A total of 14 controlling factors were selected, which included those associated with the disaster-inducing factors (P) and disaster-breeding environmental factors (DEM, SL, PlC, HS, SA, Q_{ACC}, DTR, VFD, LUP, LG, TWI, STI, $NDVI$). All three IAI models attained accurate results to generate flash flood susceptibility maps. However, LGBoost outperformed RF and CatBoost. In a similar work, 13 controlling factors, including a disaster-inducing factor (P) and disaster-breeding environmental factors (AL, SL, SA, PlC, PrC, $DisR$, $DisRd$, LUP, LG, $SDep$, SPI, TWI) were used to identify areas prone to flash flooding using ERT and different variants of RF [119]. Using 256 flood susceptibility points and 256 randomly chosen points in a watershed, and allocating 70% of the data to model training, the authors concluded that ERT showed better prediction accuracy than RF. Although the authors performed collinearity analysis to determine linear dependency among the predictors to avoid redundancy, such analysis is not required for ERT and RF. The authors concluded that topographical and hydrological features are the most critical features in flood flash predictions. Such feature importance analyses can alternatively be performed using SHAP analysis, which can also unfold interrelations and interdependencies among the predictors.

AI models have also been used for regional-scale flood hazard risks. The RF model was used for regional-scale categorical flood hazard risk assessments over 27,363 km^2 with 5000 sample points in the Dongjiang River Basin in China [120]. The predictors included disaster-inducing factors ($M3DP$, TF, RD) and disaster-breeding environmental factors (SL, DEM, DTR, $NDVI$, LUP, ST, TWI, SPI). The authors considered four risk levels, including highest (with the shortest recurrence interval), high, low, and lowest (with the longest recurrence interval) based on historical flood data. They compared predictive accuracy of RF against SVM and noted that both models identified regions with different flood risks reasonably well. Moreover, based on the Gini index, $M3PD$, RD, TF, DEM, and TWI were the most critical factors to assess flood hazard risks. Similarly, the optimized GBoost, XGBoost, RF, SVM, MLP, and Convolutional Neural Network (CNN) were used to develop a flood risk map to identify regions with low, moderate, high, and highest risk in the Pearl River Delta in China, based on information obtained from flood risk inventory maps [121]. Different from [120,121] also included disaster-bearing body factors in the AI-based decisions. Using GBoost, XGBoost, RF, SVM, MLP, and CNN, the authors evaluated flood risk using disaster-inducing factors ($M3HP$, $M1DP$, $DPE25$, TF), disaster-breeding environmental factors (DEM, SL, DTR, RD, TWI, CN), and disaster-bearing body factors (PD, $GDPD$). They used the split ratio of 70:30 for the training and test datasets. Predictive accuracy of the AI models was reported to be in the order of GBoost > XGBoost > RF∼CNN > MLP > SVM, in which, flood risk prediction accuracy of GBoost, XGBoost, and RF (IAI models) outperformed CNN, MLP, and SVM (non-IAI models). Based on the Gini index analysis of the GBoost predictions, the authors concluded that DEM, $M1DP$, RD, $DPE25$, and $M3HP$ were the most critical predictors in the order of importance for flood risk assessments. Validation of these findings and their extension to urban, rural, and coastal areas under different climate zones using XAI models using SHAP analysis are worth investigating further in follow-up studies. As for the flash flood risk assessments, the XGBoost and Least Square Support Vector Machine (LLSVM) models were used to develop flash flood risk maps for the 390,000 km^2 study area in China [122]. The authors assessed the flood risks based on information on disaster-inducing factors (annual M3HP and M31D, annual P), disaster-breeding environmental factors (DEM, SL, RD, VC, CN, TWI, SM), disaster-bearing body factors (POP, GDP), and flash flood prediction efforts. Their training data included both flash-flooded and randomly selected non-flooded sites, and allocated 70% of the data for model training. They concluded that XGBoost (an IAI model) outperformed LLSVM (a non-IAI model) in predicting the flash flood risk. Although the authors noted that XGBoost cannot provide factor importance analysis after model development, XGBoost can indeed perform such analysis when it is coupled with the SHAP method, as demonstrated in [17,18].

As originally noted in [120], neither of these AI-based flood prediction models can address the influence of hydraulic mitigation structures (e.g., dikes, levees, reservoirs) that play an important role in flood control and reduce the associated risk. Interventional AI modeling could be the proper method for such analysis in the near future.

4.2.10. Drought Predictions

Integration of drought predictions into societal decision-making processes are critical for sustainable and climate-resilience water, irrigation, and ecohydrology managements [123]. Predictors used for IAI-based drought predictions in recent studies are summarized in Table 8.

Table 8. Factors and predictors used in IAI-based drought predictions.

Factors	Predictors
Meteorologic	Precipitation (P), Temperature (T_a), Minimum temperature (T_{min}), Maximum temperature (T_{max}), Relative humidity (RH), Wind speed (U_w), Atmospheric pressure (P_a)
Climatic	Pacific decadal oscillation (PDO), Southern oscillation index (SOI), Interdecadal Pacific oscillation (IDO), Atlantic multidecadal oscillation (ADO), North Atlantic oscillation (NOA), and Oceanic Niño index (ONI)
Hydro-climatic and soil-associated	Actual evapotranspiration (ET_a), Normalized difference vegetation index ($NDVI$), Land surface temperature (LST), Soil moisture (SM)
Surface water	Surface water discharge (Q_s), Surface water temperature (T_{SW}), Surface water level (SWL)

The performance of optimized Decision Trees, AdaBoost, RF, and ERT was compared against the MLR in predicting hydrological droughts in ungauged areas in two watersheds in South Korea using remotely sensed data from six other watersheds [124]. The authors used 16 years of monthly data acquired multiple locations from 2002 to 2017 and allocated ~70% of the data to model training. Drought severity was expressed at the 3-, 6-, 9-, and 12-month time scales in terms of monthly streamflow percentiles and related to meteorologic (monthly P) and hydroclimatic and soil-associated (ET_a, $NDVI$, LST, SM) factors, in addition to the month of the year. The study concluded that AdaBoost (with the best prediction accuracy), RF, and ERT (IAI models) successfully detected observed hydrological droughts. The authors used permutation importance scores to identify the order of importance of the predictors. The analysis revealed that P, followed by SM (at the 3-month time scale) or $NDVI$ (at longer time scales) are the most critical predictors in forecasting hydrological droughts. As the authors noted, this IAI framework can be used to predict hydrological droughts in ungauged watersheds, if the ungauged basin characteristics are similar to gauged basins used in model training, suggesting that such applications require a priori domain knowledge.

The XGBoost and ANN models were used for drought forecasts based on the Standardized Precipitation Evapotranspiration Index (SPEI) 1–6 months in advance. The authors used AI models to predict SPEI in a study area in the northwest part of China from monthly-averaged meteorological and climatic variables, their lagged relationships including SPEI, and month of the year [125]. The meteorological variables included P_a, T_a, T_{min},T_{max}, RH, U_w, P, and sunshine duration using data from 32 stations during 1961 to 2016. They computed the ET_o through the PME. Climate predictors involved PDO, SOI, IDO, ADO, NOA, and ONI. They used sunshine duration as a surrogate variable for R_a, as R_s measurements were not available at the stations. They concluded that XGBoost (an IAI model) outperformed ANN (a non-IAI model) for overall droughts and drought categories. The author used a distributed lag nonlinear model to select the optimal predictors and their lag time; however, they did not disclose the order of importance of the predictors and their dependency relations, which could have been revealed by XGBoost-SHAP (a XAI model). The authors used linear booster with the XGBoost model, and noted that prediction accuracy could have improved if tree-booster was implemented instead.

Different from index-based drought predictions, the performance of the optimized RF, DT, and LSTM models were compared in predicting SWL, Q_s, T_{SW}, and GWL in low flow periods, corresponding to drought events, as well as for the entire monitoring period across the Netherlands [126]. The predictors included daily P, ET, Q_s associated with the main rivers feeding the river system of the country, sea level, and their first three lags with or without water management decisions during previous droughts, accounted for by reconstructed historical Q_s of the main water infrastructures. Using 60% of the data acquired from ∼4000 stations between 1980 and 2019 for model training, RF provided the best overall accuracy. The AI models reportedly resulted in acceptable predictions for Q_s, SWL, and T_{SW}, but relatively less prediction accuracy for GWL. Although predictors associated with the water management decisions did not improve prediction accuracy more than 9%, they appeared to be critical features at some locations. The authors tried to predict SWL and Q_s in low flow periods, at which RF and LSTM performed better, yet the predicted SWL and Q_s were 15–20% and 5–12% lower than observed values and did not reliably capture the prolonged 2018 drought. Although the authors called the AI models in their study the black-box models, the DT and RF models are not black-box models [10], as these model are amenable to coupled with the SHAP and LIME methods (forming XAI models) that can unveil the interpretable relationships between predictors and predictand, explainable model decisions, and seek new knowledge, as discussed in Section 2. Moreover, the authors used model coefficients from statistical models (e.g., LASSO) to determine which predictors have an inverse relationship with the predictors. However, such information can be readily and accurately be obtained using RF-SHAP (a XAI model) without resorting to statistical models [17,18].

4.2.11. Climate Change Impacts Modeling

Global circulation models (GCMs) that simulate physical processes in the atmosphere, ocean, cryosphere, and land surface are the primary tools to generate climate forecasts. When compared with surface observations, these models, however, suffer from biases and are unable to provide ready-to-use information at the regional spatial scales. Therefore, downscaling methods are commonly used to link the coarse-resolution global simulated predictors to the local observed predictand over the area of interest [127]. IAI and XAI models have been recently used to develop procedures for multi-model ensemble climate simulations and forecasting hydroclimatic variables under future climate scenarios.

An optimized RF model was used to develop a procedure for multi-model ensemble climate simulations from 24 Coupled Model Intercomparison Project Phase 6 (CMIP6) models to capture the characteristics of the spatially varying observed climatic data across China [128]. Each CMIP6 model was treated as a feature in the RF framework. The split ratio for the training and testing data was ∼60:40, and the length of the training data was 31,552. The predictors of the IAI model included T_a, annual T_{max}, annual T_{min}, total P in wet days, annual maximum consecutive 5-day P amount, and annual total P for events exceeding the 95th percentile. The authors reported that RF exhibited higher predictive accuracy than LR and simple arithmetic mean. They subsequently used the trained RF model to predict the regional projection of future climate for 1.5 °C, 2 °C and 3 °C global warming targets, relative to preindustrial levels, under the SSP5 emission scenario. SSP5 is the worst-case climate scenario, in which the future presumably heavily relies on intense use of fossil fuels without implementing sound adaptation and mitigation strategies. Although CMIP6 models were used as features in their RF model, the relative conformity (ranking) of the CMIP6 models to the observed data was not disclosed. This could have been effectively implemented with an RF-SHAP approach (an XAI model). We expect that the order of conformity of the CMIP6 models would vary with geographic regions and climatic zones. Therefore, it would be useful to know which CMIP6 models would be more representative for certain geographic regions and climatic zones across the globe.

As for predicting hydroclimatic variables under potential future climates, the optimized RF model was used to predict potential changes in water regime types in the northwest of the European part of Russia for the period of 2087–2099 using projected monthly runoff data from GCMs [129]. The authors divided the study area into uniform grids with the spatial resolution of 0.5° × 0.5°. They reanalyzed and computed historical monthly runoffs using the GR4J hydrological model [130] at each grid cell, which furnish the predictors for the IAI model. The RF model was trained using historical data, including the categorical water regime types as the predicant and monthly runoffs as the predictors. The authors used four GCMs, including GFDL-ESM2M, HadGEM2-ES, IPSL-CM5A-LR, and MIROC5, with three representative concentration pathways (RCP)- RCP 2.6, RCP 6.0, and RCP 8.5, to estimate future projected monthly runoffs. Here, RCP 2.6 represents the future with widely used renewable green energy, while RCP 8.5 represents the future with intense uses of oil and gas for energy production. RF was used to predict the spatial distribution of water regime types across the study area using monthly runoff computed by the GR4J model using projected climate data from the GCM models under different RCP scenarios. The analysis suggested that water regimes types could alter over 73.6% and 99% of the study area under the RCP 2.6 and RCP 8.5 scenarios, respectively during the 2087–2099 period. Moreover, the summer and winter flows could be less stable and spring flow peaks could be lower while shifting to earlier times. Although the authors used historical and projected climate data in calculating monthly runoff using the hydrological model, climate variables could have been also used as predictors in the IAI model. In this case, interdependencies and the importance of the predictors as well as their critical values responsible for changes in water regime types could have been determined by using the RF-SHAP model (i.e., XAI model).

Moreover, a novel optimized XGBoost-based XAI framework to predict long-term GWL and decadal hydrological droughts in an ecologically fragile groundwater-dependent semi-arid region in south-central Texas, USA under projected future climate scenarios from 2021 to 2100 was presented in [20]. The severity of future hydrological droughts was assessed based on mandated groundwater pumping reductions, if the tiered critical period management pumping restriction plan as part of the current habitat conservation measures at the site, would have been implemented during the seven years-long worst drought that the study region experienced in 1950s. Groundwater pumping reductions in this plan hinge on GWL at an index well. The authors set-up the XAI model first to predict weekly GWL from a set of weekly features, including historical lagged GWL, lagged and current P, and current T_{min} and T_{max}. They used the recorded weekly climate data from 1950 to 2005 to train the XGBoost model. When combined with the SHAP method, the trained XGBoost model revealed that the first lag of GWL and P, in addition to T_{max} were the most decisive features to predict GWL. The trained XAI predicted GWL from 2006 to 2020 with high accuracy when historical climate data or Coupled Model Intercomparison Project Phase 5 (CMIP5) data under the RCP 4.5 and 8.5 scenarios were used as input. In their study, CMIP5 data were downscaled using the Multivariate Adaptive Constructed Analogs (MACA) [131]. Subsequently, the validated XGBoost model was used with the CMIP5-MACA projected T_{max} and P to forecast weekly GWL and decadal hydrological droughts from 2021 to 2100 under the RCP 4.5 and 8.5 scenarios. The XAI model additionally revealed that despite an increasing precipitation trend, compound effects of increased evapotranspiration, lower soil moisture, and reduced diffuse recharge due to warmer temperatures could amplify severe hydrological droughts that lower groundwater levels, if regional-scale climate adaptation and mitigation strategies are not implemented.

5. Discussion and Conclusions

The review identified several important implications that need to be considered in IAI/XAI models:

Explainability of the IAI-predicted results: Explanatory methods such as SHAP and LIME could enhance the accountability and trustworthiness of the IAI-based inferences and decisions in practice. IAI models can be trusted and used more often, if they (i) can explain the reasoning behind the AI-based decisions, (ii) unveil how the decision can be further enhanced using information on the order of importance of features while considering their complex and nonlinear interdependencies and interrelations, and (iii) are amenable to set up testable hypotheses and probabilistic analysis to unveil favorable conditions for enhanced targeted decisions, as demonstrated in Ref [18].

Multiple IAI/XAI models in decision-making: The Rashomon set argument [132] implies that if the data permits a large set of reasonably accurate AI models to exist, this large set of accurate models often contains at least one AI model that is interpretable (IAI). This model is thus both interpretable and accurate [26], which is imperative for the explainability and scientific value of the outcome [7]. Moreover, diverse IAI/XAI models could perform differently on distinct hydroclimatic problems, as well as on the same type of problems but with different sets of site-specific predictors of distinct lengths and types. Therefore, multiple IAI/XAI models should be used in practice to identify the problem- and site-specific best-performing IAI/XAI model(s) with the highest prediction precision while bounding prediction uncertainties.

Spatial Scale in IAI/XAI-based analysis: IAI models have been used to analyze and predict hydroclimatic processes from a watershed-scale [96] to a continent-scale [101], as long as sufficient, high quality data are available to train the models. Advances in remote sensing, data acquisition, and data analysis tools allow multiscale applications of the IAI/XAI modeling.

Domain knowledge: Prediction of certain hydroclimatic processes using data-driven IAI/XAI modeling at particular sites with scarce measurements would require strong domain knowledge. For example, information about aquifer properties and groundwater levels could be scarce, yet the knowledge on groundwater potential could be imperative for further development and management of water resources at particular sites. In such circumstances, if the groundwater potential is known to be controlled by easy-to-access topographic and geologic features and/or related to springs inventory, groundwater potential could still be predicted using IAI models even in the absence of detailed hydrogeologic data, as shown in Refs. [72,73].

Balanced data in categorical decisions: Imbalanced classes in categorical IAI/XAI analysis (e.g., identifying severe flood-risk regions, high groundwater potential sites) need a comprehensive analysis using confusion matrices (unveiling false positive and false negative) and it may require one to balance the imbalance classes, as in [33] for more accurate and robust predictions. Unbalanced data could result in biased and unreliable predictions.

Hybrid IAI/XAI and non-IAI modeling: Several recent studies (e.g., Refs. [34,81]) used IAI and non-IAI models to enhance the accuracy in AI-based decisions. In such hybrid modeling, the IAI model is commonly used to identify the most influential predictors on the decision and unveil the nonlinear correlative effect between the predictors and predictands. This information is then fed into the non-IAI model, which is ultimately used as a predictive tool. This approach, however, could induce a risk for proper training of a non-IAI model as it uses a reduced predictor list determined by the IAI model. This could diminish the predictive accuracy of the non-IAI model, as the underlying algorithms, mechanisms, and assumptions of the IAI and non-IAI models are different.

Interventional modeling: The IAI/XAI models make predictions based on historical events and data. These models will not be able to predict unprecedented events, as they would not have any a priori knowledge about such events and the associated nonlinear correlative relations between the predictors and predictands. This implies that traditional non-interventional IAI/XAI modeling hinges on stationary assumption, in which the historical statistical predictors–predictands relations would presumably be valid in the future. However, such stationary assumptions may not be valid in hydroclimatic domains under intensifying human inferences and future climates. Therefore, when the IAI model is used for prediction and if an unprecedented event were to occur during the prediction interval, the IAI model analysis can be intervened and the IAI model is re-trained using the new information about the first-time occurring event, as is implemented in [83]. We envision that the interventional AI modeling will be useful in practice, as hydroclimatic systems continue to be altered by human impacts [133] and climate change [134], which have been already affecting the frequency, intensity, and magnitude of the extreme events.

IAI/XAI modeling vs. physics-based modeling: IAI/XAI models have emerged as a reliable simulator and predictive tool as an alternative or complementary to physics-based models. Unlike the latter, the IAI models do not require any assumptions on the system dynamics or a set of governing equations to accurately represent the types of physical mechanisms (e.g., flow) in different parts of the domain. They can be applied using easy-to-access meteorologic, topographic, and satellite-derived data for hydroclimatic predictions over spatially-heterogeneous sites, as shown in Refs. [67,70,72]. In some problems, the IAI models exhibited better or comparable performance to physics-based models [67,104]. In other problems, physics-based and IAI models were coupled to enhance the prediction accuracy [113]. At sites with limited time-variant data, physics-based models have been used to generate additional synthetic data to train AI models [17,18].

IAI/XAI modeling in citizen science projects based on crowdsourcing: Crowdsourced distributed hydrologic measurements contributed by the public (e.g., using a smartphone app) have been considered as a potential supplement for data networks in hydrological research [135]. Although this could help fill the data gap, uncertainty and error in citizen science measurements are a primary concern for the scientific community [136]. Thus, a decision tree model was recently used as a quality control filter to flag potentially erroneous data points in citizen science data of the stream stage [137]. The decision tree model can also be used with different sources of datasets (e.g., precipitation, water quality) to determine the ruleset for the incorrect and atypical values. We expect to see the use of IAI/XAI models in establishing various problem-specific data quality controls in nonsystematically acquired large citizen science measurements to flag suspicious data in an effort to reduce false positives and false negatives in confusion matrices.

XAI modeling and decisions on the fly: Big data and predictive analytics can potentially provide accurate, real-time or near real-time analytics and insights in real-life hydroclimatic applications involving prediction of recurrence and impacts of natural hazards such as floods, droughts, soil erosion, and development of mitigation measures to reduce their adverse impacts [3]. In the near future, we expect to see that XAI models with new online tools would be used to make prediction or decisions on the fly as new data are streamed in. In such applications, explainability of the underlying reasoning of AI-based decision would be paramount for the stakeholders for the enhanced reliability, trustworthiness, accountability of the decisions, and development of timely and effective mitigation and adaptation measures.

Automated XAI modeling: MLMapper, which operates with 20 supervised AI models, was recently used to map the surface area of wetlands from geological, geomorphological, hydrogeological, and biological data [93]. In the near future, we expect that an automated XAI modeling framework involving multiple AI methods be widely used for prediction and future projection of hydroclimatic processes. Such an encapsulated framework would automatically deploy AI models from a suite of AI models and select the AI models with the highest prediction accuracy on the test data. The framework would call the explanatory methods (e.g., SHAP and LIME methods) to determine local and global analyses to identify the most critical features by considering interdependencies and interrelations among the predictors, remove the least critical features from the predictor list, and re-engineer, optimize, and transfer the selected IAI model to XAI models with enhanced accountability. Those XAI models would then be used for scenario-based future projections and construction of testable hypotheses to identify the conditions at which the projected predictand could further be enhanced, based on a specific combination of predictors. Such frameworks would increase the trustworthiness and fidelity of the IAI/XAI models.

A generalized XAI framework for a hydroclimatic application is shown in Figure 2. In this framework, Step (I) involves data acquisition, quality checks, curation, and data imputation when necessary. The data could include static, time-series, numerical, categorical, point, and gridded data. The data may also include externally-acquired projected predictors (e.g., future climate variables downscaled from global circulation models) if the goal is to project future target variables. In Steps (II) and (III), grid search hyperparameter optimization could be implemented to determine the optimal set of parameters for the chosen AI models using the training data. Multi-fold cross-validation techniques are typically used to tune the models and determine the predictive accuracy of the AI model on the unseen test data. If the prediction performance of the model is found to be statistically significant, the optimized AI model can then be used as a predictor tool in Step IV. At this point, the AI model can be used to predict future values of predictands (e.g., groundwater levels, hydrologic droughts).

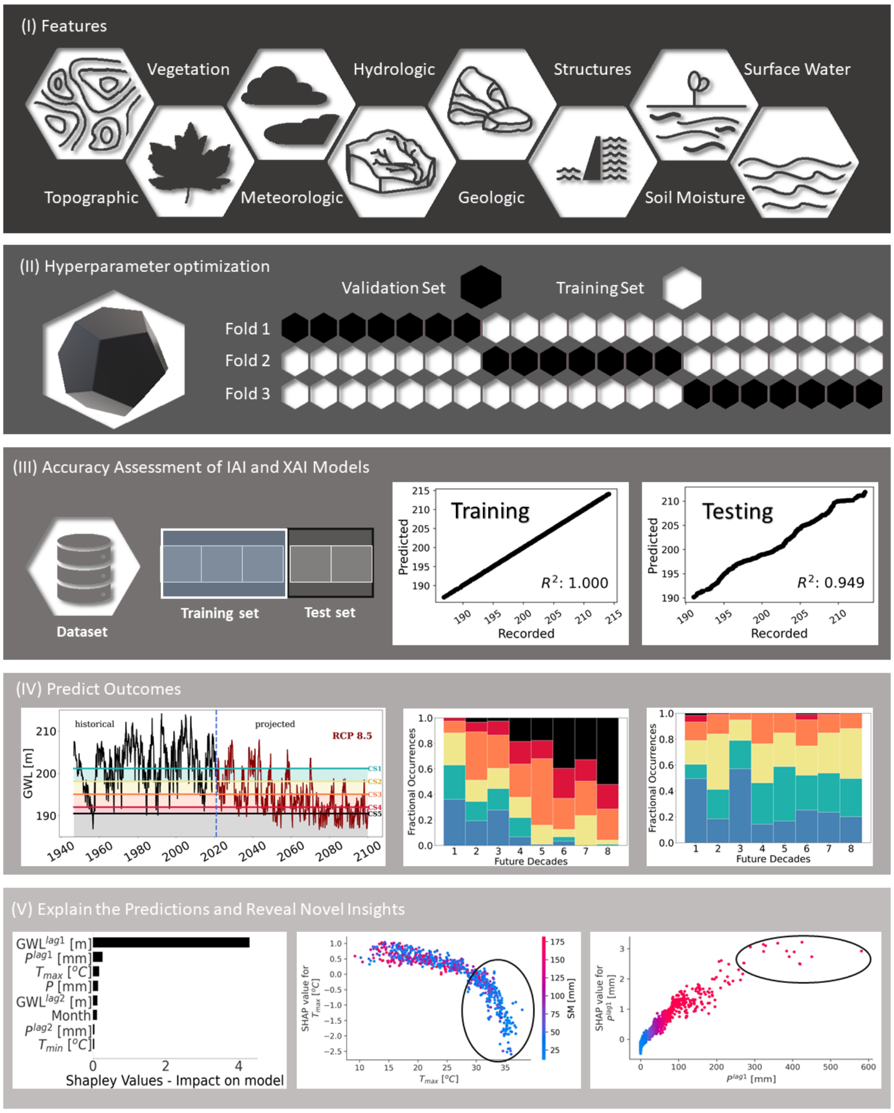

Figure 2. A generalized framework of an XAI model. Basic steps include: (**I**) collection and curation of hydroclimatic data sets to be used to train the AI model and test its prediction performance; (**II**) grid search hyperparameter optimization using multi-fold cross validations to find the best set of parameters for AI model runs, (**III**) AI model training and determination of its prediction accuracy on test data based on statistical measures; (**IV**) prediction/projection of hydroclimatic variables using the validated AI model; and (**V**) determination of the rank of the most influential predictors in predicting target variables based on global SHAP analyses (**left** panel), and determination of the inflection point of the predictors (where the SHAP values on the y-axis change signs), based on the local SHAP analysis, above or below which the value of the predictand would increase or decrease (**right** panel). The predictive AI model is transitioned into the IAI and XAI models in (**V**).

The tree-based ensemble AI models are interpretable AI models as they are amenable to be fused with the explanatory methods to unveil the nonlinear correlative relations between the predictors and predictands in model outputs in Step (IV). The AI model turns into an explainable model in Step (V) when it is combined with explanatory methods (e.g., SHAP). The global SHAP analysis in Step (V) identifies the order of importance of the predictors by explicitly accounting for their interrelations and interdependencies. At this point, the user can implement 'feature engineering' to reduce the size of the input data set (i.e, number of predictors) by selecting only the topmost influential features and repeat the Steps (III) and (IV). The local SHAP analysis in Step (V) identifies the inflection points of predictors above and below which the predictands would further increase and decrease. This step is critical to unveil new information (e.g., a critical T_a above which soil moisture and diffused recharge diminish) and establish testable hypotheses how the predictand may vary probabilistically if the related hydrologic conditions vary under future conditions, as shown in Refs. [17,18,20]. These are key analyses to peek into the internal logic of the AI modeling and enhance the accountability and trustworthiness of AI-based decisions in practice.

Author Contributions: H.B. and D.C. formulated the concept and structured the paper. All authors were involved in the conceptualization & literature review and wrote, reviewed, and edited the paper. All authors have read and agreed to the published version of the paper.

Funding: This research received external funding from the Edwards Aquifer Authority (award numbers SAT0003036 and SAT0003314), the U.S. Geological Survey's South Central Climate Adaptation Science Center (award number G21AC10751), and the U.S. Department of Agriculture (award number SAT0003387).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This material is based upon work supported by the Edwards Aquifer Authority, the U.S. Geological Survey's South Central Climate Adaptation Science Center, and the U.S. Department of Agriculture. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the Edwards Aquifer Authority or the U.S. Geological Survey or the U.S. Department of Agriculture.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

Commonly used abbreviations in the paper for Artificial Intelligence models and key variables:

Artificial Intelligence Models:

AdaBoost	Adaptive Boosting
ANN	Artificial Neural Networks
CNN	Convolutional Neural Network
DL	Deep Learning
DT	Decision Trees
GBoost	Gradient Boosting
IAI	Interpretable Artificial Intelligence
LIME	Local Interpretable Model-agnostic Explanations
LSTM	Long Short Term Memory
LR	Linear Regression
AI	Artificial Intelligence
MLP	Multi-layer Perceptron
XGBoost	Natural Gradient Boosting
RF	Random forest
SHAP	SHaply Additive Explanation
SVM	Support Vector Machine

SVR	Support Vector Regression
XGBoost	Extreme Gradient Boosting
XAI	Explainable Artificial Intelligence
Key Variables:	
ET	Evapotranspiration
ET_a	Actual evapotranspiration
ET_o	Reference crop evapotranspiration
GWL	Groundwater level
GWP	Groundwater potential
P	Precipitation
P_a	Atmospheric pressure
Q_s	Streamflow
R_s	Shortwave solar radiation
SM	Soil moisture
T_a	Air temperature
T_{sw}	Water surface temperature
U_w	Wind speed
W_L	Water level

References

1. Ochoa-Tocachi, B.F.; Buytaert, W.; Antiporta, J.; Acosta, L.; Bardales, J.D.; Célleri, R.; Crespo, P.; Fuentes, P.; Gil-Ríos, J.; Guallpa, M.; et al. High-resolution hydrometeorological data from a network of headwater catchments in the tropical Andes. *Sci. Data* **2018**, *5*, 180080. [CrossRef] [PubMed]
2. Singh, V.P. Hydrologic modeling: Progress and future directions. *Geosci. Lett.* **2018**, *5*, 15. [CrossRef]
3. Adamala, S. An Overview of Big Data Applications in Water Resources Engineering. *Mach. Learn. Res.* **2017**, *2*, 10–18. [CrossRef]
4. Obermeyer, Z.; Emanuel, E.J. Predicting the Future—Big Data, Machine Learning, and Clinical Medicine. *N. Engl. J. Med.* **2016**, *375*, 1216–1219. [CrossRef]
5. Biran, O.; Cotton, C.V. Explanation and Justification in Machine Learning: A Survey. IJCAI 2017 Workshop on Explainable Artificial Intelligence. 2017. Available online: http://www.cs.columbia.edu/~orb/papers/xai_survey_paper_2017.pdf (accessed on 19 February 2022).
6. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38. [CrossRef]
7. Roscher, R.; Bohn, B.; Duarte, M.F.; Garcke, J. Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access* **2020**, *8*, 42200–42216. [CrossRef]
8. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
9. Zounemat-Kermani, M.; Batelaan, O.; Fadaee, M.; Hinkelmann, R. Ensemble machine learning paradigms in hydrology: A review. *J. Hydrol.* **2021**, *598*, 126266. [CrossRef]
10. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Pedreschi, D.; Giannotti, F. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* **2018**, *51*, 93. [CrossRef]
11. Shapley, L. A value for *n*-person games. *Contrib. Theory Games* **1953**, 307–317.
12. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 2522–5839. [CrossRef] [PubMed]
13. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
14. Xie, Y.R.; Castro, D.C.; Bell, S.E.; Rubakhin, S.S.; Sweedler, J.V. Single-Cell Classification Using Mass Spectrometry through Interpretable Machine Learning. *Anal. Chem.* **2020**, *92*, 9338–9347. [CrossRef] [PubMed]
15. Rodríguez-Pérez, R.; Bajorath, J. Interpretation of machine learning models using shapley values: Application to compound potency and multi-target activity predictions. *J. Comput. Aided Mol. Des.* **2020**, *34*, 1013–1026. [CrossRef] [PubMed]
16. Mangalathu, S.; Hwang, S.H.; Jeon, J.S. Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach. *Eng. Struct.* **2020**, *219*, 110927. [CrossRef]
17. Başağaoğlu, H.; Chakraborty, D.; Winterle, J. Reliable Evapotranspiration Predictions with a Probabilistic Machine Learning Framework. *Water* **2021**, *13*, 557. [CrossRef]
18. Chakraborty, D.; Başağaoğlu, H.; Winterle, J. Interpretable vs. noninterpretable machine learning models for data-driven hydro-climatological process modeling. *Expert Syst. Appl.* **2021**, *170*, 114498. [CrossRef]

19. Chakraborty, D.; Ivan, C.; Amero, P.; Khan, M.; Rodriguez-Aguayo, C.; Başağaoğlu, H.; Lopez-Berestein, G. Explainable Artificial Intelligence Reveals Novel Insight into Tumor Microenvironment Conditions Linked with Better Prognosis in Patients with Breast Cancer. *Cancers* **2021**, *13*, 3450. [CrossRef]
20. Chakraborty, D.; Başağaoğlu, H.; Gutierrez, L.; Mirchi, A. Explainable AI reveals new hydroclimatic insights for ecosystem-centric groundwater management. *Environ. Res. Lett.* **2021**, *16*, 114024. [CrossRef]
21. Chakraborty, D.; Alam, A.; Chaudhuri, S.; Başağaoğlu, H.; Sulbaran, T.; Langar, S. Scenario-based prediction of climate change impacts on building cooling energy consumption with explainable artificial intelligence. *Appl. Energy* **2021**, *291*, 116807. [CrossRef]
22. Li, L.; Qiao, J.; Yu, G.; Wang, L.; Li, H.Y.; Liao, C.; Zhu, Z. Interpretable tree-based ensemble model for predicting beach water quality. *Water Res.* **2022**, *211*, 118078. [CrossRef]
23. Wang, D.; Thunéll, S.; Lindberg, U.; Jiang, L.; Trygg, J.; Tysklind, M. Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based machine learning methods. *J. Environ. Manag.* **2022**, *301*, 113941. [CrossRef] [PubMed]
24. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2021**, *23*, 18. [CrossRef] [PubMed]
25. Lipton, Z.C. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. *Queue* **2018**, *16*, 31–57. [CrossRef]
26. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef]
27. Eschenbach, W.J. Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philos. Technol.* **2021**, *34*, 1607–1622. [CrossRef]
28. D'Isanto, A.; Cavuoti, S.; Gieseke, F.; Polsterer, K.L. Return of the features—Efficient feature selection and interpretation for photometric redshifts. *Astron. Astrophys.* **2018**, *616*, A97. [CrossRef]
29. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLoS ONE* **2018**, *13*, e0194889. [CrossRef]
30. Shin, D. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *Int. J. Hum.-Comput. Stud.* **2021**, *146*, 102551. [CrossRef]
31. Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V.I. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 310. [CrossRef]
32. London, A.J. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Cent. Rep.* **2019**, *49*, 15–21. [CrossRef]
33. Bedi, S.; Samal, A.; Ray, C.; Snow, D. Comparative evaluation of machine learning models for groundwater quality assessment. *Environ. Monit. Assess.* **2020**, *192*, 776. [CrossRef] [PubMed]
34. Ravindran, S.; Bhaskaran, S.; Ambat, S. A Deep Neural Network Architecture to Model Reference Evapotranspiration Using a Single Input Meteorological Parameter. *Environ. Process* **2021**, *103*, 1567–1599. [CrossRef]
35. Wen, Y.; Zhao, J.; Zhu, G.; Xu, R.; Yang, J. Evaluation of the RF-Based Downscaled SMAP and SMOS Products Using Multi-Source Data over an Alpine Mountains Basin, Northwest China. *Water* **2021**, *13*, 2875. [CrossRef]
36. Ottenhoff, M.C.; Ramos, L.A.; Potters, W.; Janssen, M.L.F.; Hubers, D.; Hu, S.; Fridgeirsson, E.A.; Piña-Fuentes, D.; Thomas, R.; van der Horst, I.C.C.; et al. Predicting mortality of individual patients with COVID-19: A multicentre Dutch cohort. *BMJ Open* **2021**, *11*, e047347. [CrossRef] [PubMed]
37. Ben Jabeur, S.; Khalfaoui, R.; Ben Arfi, W. The effect of green energy, global environmental indexes, and stock markets in predicting oil price crashes: Evidence from explainable machine learning. *J. Environ. Manag.* **2021**, *298*, 113511. [CrossRef] [PubMed]
38. Zhang, W.; Zhang, R.; Wu, C.; Goh, A.T.C.; Lacasse, S.; Liu, Z.; Liu, H. State-of-the-art review of soft computing applications in underground excavations. *Geosci. Front.* **2020**, *11*, 1095–1106. [CrossRef]
39. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
40. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*; NIPS: Long Beach, CA, USA, 4–9 December 2017.
41. Dorogush, A.V.; Ershov, V.; Gulin, A. CatBoost: Gradient boosting with categorical features support. *arXiv* **2018**, arXiv:1810.11363.
42. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [CrossRef]
43. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
44. Little, J.L.; Rubin, D.A. *Statistical Analysis with Missing Data*; John Wiley: New York, NY, USA, 1987.
45. oi: 10.1029/2006WR005298 Gill, M.K.; Asefa, T.; Kaheil, Y.; McKee, M. Effect of missing data on performance of learning algorithms for hydrologic predictions: Implications to an imputation technique. *Water Resour. Res.* **2007**, *43*. [CrossRef]
46. Teegavarapu, R.S.V. Statistical corrections of spatially interpolated missing precipitation data estimates. *Hydrol. Process.* **2014**, *28*, 3789–3808. [CrossRef]
47. Miró, J.J.; Caselles, V.; Estrela, M.J. Multiple imputation of rainfall missing data in the Iberian Mediterranean context. *Atmos. Res.* **2017**, *197*, 313–330. [CrossRef]

48. Aguilera, H.; Guardiola-Albert, C.; Serrano-Hidalgo, C. Estimating extremely large amounts of missing precipitation data. *J. Hydroinform.* **2020**, *22*, 578–592. [CrossRef]
49. Stekhoven, D.J.; Bühlmann, P. MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics* **2012**, *28*, 112–118. [CrossRef]
50. Arriagada, P.; Karelovic, B.; Link, O. Automatic gap-filling of daily streamflow time series in data-scarce regions using a machine learning algorithm. *J. Hydrol.* **2021**, *598*, 126454. [CrossRef]
51. Tao, X.E.; Chen, H.; Xu, C.Y.; Hou, Y.K.; Jie, M.X. Analysis and prediction of reference evapotranspiration with climate change in Xiangjiang River Basin, China. *Water Sci. Eng.* **2015**, *8*, 273–281. [CrossRef]
52. Peña-Arancibia, J.L.; Mainuddin, M.; Kirby, J.M.; Chiew, F.H.; McVicar, T.R.; Vaze, J. Assessing irrigated agriculture's surface water and groundwater consumption by combining satellite remote sensing and hydrologic modelling. *Sci. Total Environ.* **2016**, *542*, 372–382. [CrossRef]
53. Allen, R.G.; Pereira, L.S.; Raes, D.; Smith, M. *Crop Evapotranspiration–Guidelines for Computing Crop Water Requirements*; FAO Irrigation and Drainage Paper 56; FAO: Rome, Italy, 1998; ISBN 92-5-104219-5.
54. Wu, L.; Fan, J. Comparison of neuron-based, kernel-based, tree-based and curve based machine learning models for predicting daily reference evapotranspiration. *PLoS ONE* **2019**, *14*, e0217520. [CrossRef]
55. Zhang, Y.; Zhao, Z.; Zheng, J. CatBoost: A new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China. *J. Hydrol.* **2020**, *588*, 125087. [CrossRef]
56. Huang, G.; Wu, L.; Ma, X.; Zhang, W.; Fan, J.; Yu, X.; Zeng, W.; Zhou, H. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. *J. Hydrol.* **2019**, *574*, 1029–1041. [CrossRef]
57. Tang, D.; Feng, Y.; Gong, D.; Hao, W.; Cui, N. Evaluation of artificial intelligence models for actual crop evapotranspiration modeling in mulched and non-mulched maize croplands. *Comp. Electron. Agric.* **2018**, *152*, 375–384. [CrossRef]
58. Sun, Q.; Miao, C.; Duan, Q.; Ashouri, H.; Sorooshian, S.; Hsu, K.L. A Review of Global Precipitation Data Sets: Data Sources, Estimation, and Intercomparisons. *Rev. Geophys.* **2018**, *56*, 79–107. [CrossRef]
59. Tian, C.; Wang, L.; Kaseke, K.; Broxton, W.B. Stable isotope compositions $\delta^2 H$, $\delta^{18} O$ and$\delta^{17} O$) of rainfall and snowfall in the central United States. *Sci. Rep.* **2018**, *8*, 6712. [CrossRef]
60. Nelson, D.B.; Basler, D.; Kahmen, A. Precipitation isotope time series predictions from machine learning applied in Europe. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2024107118. [CrossRef]
61. Nashwan, M.S.; Shahid, S. Symmetrical uncertainty and random forest for the evaluation of gridded precipitation and temperature data. *Atmos. Res.* **2019**, *230*, 104632. [CrossRef]
62. Zhang, J.; Fan, H.; He, D.; Chen, J. Integrating precipitation zoning with random forest regression for the spatial downscaling of satellite-based precipitation: A case study of the Lancang–Mekong River basin. *Int. J. Climatol.* **2019**, *39*, 3947–3961. [CrossRef]
63. Touhami, I.; Andreu, J.; Chirino, E.; Sánchez, J.; Pulido-Bosch, A.; Martínez-Santos, P.; Moutahir, H.; Bellot, J. Comparative performance of soil water balance models in computing semi-arid aquifer recharge. *Hydrol. Sci. J.* **2014**, *59*, 193–203. [CrossRef]
64. Wagner, W.; Naeimi, V.; Scipal, K.; de Jeu, R.; Martínez-Fernández, J. Soil moisture from operational meteorological satellites. *Hydrogeol. J.* **2007**, *15*, 121–131. [CrossRef]
65. Oroza, C.A.; Bales, R.C.; Stacy, E.M.; Zheng, Z.; Glaser, S.D. Long-Term Variability of Soil Moisture in the Southern Sierra: Measurement and Prediction. *Vadose Zone J.* **2018**, *17*, 170178. [CrossRef]
66. Simunek, J.; Genuchten, M.T.V.; Sejna, M. *The HYDRUS-1D Software Package For Simulating the One-Dimensional Movement of Water, Heat, and Multiple Solutes in Variably-Saturated Media*; University of California: Riverside, CA, USA, 2005.
67. Carranza, C.; Nolet, C.; Pezij, M.; van der Ploeg, M. Root zone soil moisture estimation with Random Forest. *J. Hydrol.* **2021**, *593*, 125840. [CrossRef]
68. Nag, S.; Ghosh, P. Delineation of groundwater potential zone in Chhatna Block, Bankura District, West Bengal, India using remote sensing and GIS techniques. *Environ. Earth Sci.* **2013**, *70*, 2115–2127. [CrossRef]
69. Ahmed, N.; Hoque, M.; Pradhan, B.; Arabameri, A. Spatio-Temporal Assessment of Groundwater Potential Zone in the Drought-Prone Area of Bangladesh Using GIS-Based Bivariate Models. *Nat. Resour. Res.* **2021**, *30*, 3315–3337. [CrossRef]
70. Sachdeva, S.; Kumar, B. Comparison of gradient boosted decision trees and random forest for groundwater potential mapping in Dholpur (Rajasthan), India. *Stoch. Environ. Res. Risk Assess.* **2021**, *35*, 287–306. [CrossRef]
71. Park, S.; Kim, J. The Predictive Capability of a Novel Ensemble Tree-Based Algorithm for Assessing Groundwater Potential. *Sustainability* **2021**, *13*, 2459. [CrossRef]
72. Naghibi, S.A.; Hashemi, H.; Berndtsson, R.; Lee, S. Application of extreme gradient boosting and parallel random forest algorithms for assessing groundwater spring potential using DEM-derived factors. *J. Hydrol.* **2020**, *589*, 125197. [CrossRef]
73. Namous, M.; Hssaisoune, M.; Pradhan, B.; Lee, C.W.; Alamri, A.; Elaloui, A.; Edahbi, M.; Krimissa, S.; Eloudi, H.; Ouayah, M.; et al. Spatial Prediction of Groundwater Potentiality in Large Semi-Arid and Karstic Mountainous Region Using Machine Learning Models. *Water* **2021**, *13*, 2273. [CrossRef]
74. Eris, E.; Wittenberg, H. Estimation of baseflow and water transfer in karst catchments in Mediterranean Turkey by nonlinear recession analysis. *J. Hydrol.* **2015**, *530*, 500–507. [CrossRef]
75. Huang, F.; Huang, J.; Jiang, S.H.; Zhou, C. Prediction of groundwater levels using evidence of chaos and support vector machine. *J. Hydroinform.* **2017**, *19*, 586–606. [CrossRef]
76. Kebede, H.; Fisher, D.; Sui, R.; Reddy, K. Irrigation Methods and Scheduling in the Delta Region of Mississippi: Current Status and Strategies to Improve Irrigation Efficiency. *Am. J. Plant Sci.* **2014**, *5*, 2917–2928. [CrossRef]

77. Kleinman, P.; Spiegal, S.; Rigby, J.; Goslee, S.; Baker, J.; Bestelmeyer, B.; Boughton, R.; Bryant, R.; Cavigelli, M.; Derner, J.; et al. Advancing the Sustainability of US Agriculture through Long-Term Research. *J. Environ. Qual.* **2018**, *47*, 1412–1425. [CrossRef] [PubMed]
78. Rahman, A.S.; Hosono, T.; Quilty, J.M.; Das, J.; Basak, A. Multiscale groundwater level forecasting: Coupling new machine learning approaches with wavelet transforms. *Adv. Water Resour.* **2020**, *141*, 103595. [CrossRef]
79. Kombo, O.H.; Kumaran, S.; Sheikh, Y.H.; Bovim, A.; Jayavel, K. Long-Term Groundwater Level Prediction Model Based on Hybrid KNN-RF Technique. *Hydrology* **2020**, *7*, 59. [CrossRef]
80. Hussein, E.A.; Thron, C.; Ghaziasgar, M.; Bagula, A.; Vaccari, M. Groundwater Prediction Using Machine-Learning Tools. *Algorithms* **2020**, *13*, 300. [CrossRef]
81. Hadi, S.J.; Abba, S.I.; Sammen, S.S.; Salih, S.Q.; Al-Ansari, N.; Yaseen, Z.M. Non-Linear Input Variable Selection Approach Integrated with Non-Tuned Data Intelligence Model for Streamflow Pattern Simulation. *IEEE Access* **2019**, *7*, 141533–141548. [CrossRef]
82. Lee, C.H.; Yeh, H.F. Impact of Climate Change and Human Activities on Streamflow Variations Based on the Budyko Framework. *Water* **2019**, *11*, 2001. [CrossRef]
83. Zhang, H.; Yang, Q.; Shao, J.; Wang, G. Dynamic Streamflow Simulation via Online Gradient-Boosted Regression Tree. *J. Hydrol. Eng.* **2019**, *24*, 04019041. [CrossRef]
84. Cui, Z.; Qing, X.; Chai, H.; Yang, S.; Zhu, Y.; Wang, F. Real-time rainfall-runoff prediction using light gradient boosting machine coupled with singular spectrum analysis. *J. Hydrol.* **2021**, *603*, 127124. [CrossRef]
85. Yu, X.; Wang, Y.; Wu, L.; Chen, G.; Wang, L.; Qin, H. Comparison of support vector regression and extreme gradient boosting for decomposition-based data-driven 10-day streamflow forecasting. *J. Hydrol.* **2020**, *582*, 124293. [CrossRef]
86. Randle, T.J.; Morris, G.L.; Tullos, D.D.; Weirich, F.H.; Kondolf, G.M.; Moriasi, D.N.; Annandale, G.W.; Fripp, J.; Minear, J.T.; Wegner, D.L. Sustaining United States reservoir storage capacity: Need for a new paradigm. *J. Hydrol.* **2021**, *602*, 126686. [CrossRef]
87. Xia, R.; Zhang, Y.; Critto, A.; Wu, J.; Fan, J.; Zheng, Z.; Zhang, Y. The Potential Impacts of Climate Change Factors on Freshwater Eutrophication: Implications for Research and Countermeasures of Water Management in China. *Sustainability* **2016**, *8*, 229. [CrossRef]
88. Schulz, S.; Darehshouri, S.; Hassanzadeh, E.; Tajrishy, M.; Schüth, C. Climate change or irrigated agriculture—What drives the water level decline of Lake Urmia. *Sci. Rep.* **2020**, *10*, 236. [CrossRef] [PubMed]
89. Leibowitz, S.G.; Wigington, P.J., Jr.; Schofield, K.A.; Alexander, L.C.; Vanderhoof, M.K.; Golden, H.E. Connectivity of Streams and Wetlands to Downstream Waters: An Integrated Systems Framework. *J. Am. Water Resour. Assoc.* **2018**, *54*, 298–322. [CrossRef] [PubMed]
90. Sapitang, M.; Ridwan, W.M.; Faizal Kushiar, K.; Najah Ahmed, A.; El-Shafie, A. Machine Learning Application in Reservoir Water Level Forecasting for Sustainable Hydropower Generation Strategy. *Sustainability* **2020**, *12*, 6121. [CrossRef]
91. Guyennon, N.; Salerno, F.; Rossi, D.; Rainaldi, M.; Calizza, E.; Romano, E. Climate change and water abstraction impacts on the long-term variability of water levels in Lake Bracciano (Central Italy): A Random Forest approach. *J. Hydrol. Reg. Stud.* **2021**, *37*, 100880. [CrossRef]
92. Choi, C.; Kim, J.; Han, H.; Han, D.; Kim, H.S. Development of Water Level Prediction Models Using Machine Learning in Wetlands: A Case Study of Upo Wetland in South Korea. *Water* **2020**, *12*, 93. [CrossRef]
93. Martínez-Santos, P.; Díaz-Alcaide, S.; De la Hera-Portillo, A.; Gómez-Escalonilla, V. Mapping groundwater-dependent ecosystems by means of multi-layer supervised classification. *J. Hydrol.* **2021**, *603*, 126873. [CrossRef]
94. Cosgrove, W.J.; Loucks, D.P. Water management: Current and future challenges and research directions. *Water Resour. Res.* **2015**, *51*, 4823–4839. [CrossRef]
95. Lumb, A.; Sharma, T.; Bibeault, J. A Review of Genesis and Evolution of Water Quality Index (WQI) and Some Future Directions. *J. Environ. Chem. Eng.* **2011**, *3*, 11–24. [CrossRef]
96. Singha, S.; Pasupuleti, S.; Singha, S.S.; Singh, R.; Kumar, S. Prediction of groundwater quality using efficient machine learning technique. *Chemosphere* **2021**, *276*, 130265. [CrossRef]
97. Sahour, H.; Gholami, V.; Vazifedan, M. A comparative analysis of statistical and machine learning techniques for mapping the spatial distribution of groundwater salinity in a coastal aquifer. *J. Hydrol.* **2020**, *591*, 125321. [CrossRef]
98. Tran, D.A.; Tsujimura, M.; Ha, N.T.; Nguyen, V.T.; Binh, D.V.; Dang, T.D.; Doan, Q.V.; Bui, D.T.; Anh Ngoc, T.; Phu, L.V.; et al. Evaluating the predictive power of different machine learning algorithms for groundwater salinity prediction of multi-layer coastal aquifers in the Mekong Delta, Vietnam. *Ecol. Indic.* **2021**, *127*, 107790. [CrossRef]
99. Kumar, P.; Bansod, B.K.; Debnath, S.K.; Thakur, P.K.; Ghanshyam, C. Index-based groundwater vulnerability mapping models using hydrogeological settings: A critical evaluation. *Environ. Impact Assess. Rev.* **2015**, *51*, 38–49. [CrossRef]
100. Barzegar, R.; Razzagh, S.; Quilty, J.; Adamowski, J.; Kheyrollah Pour, H.; Booij, M.J. Improving GALDIT-based groundwater vulnerability predictive mapping using coupled resampling algorithms and machine learning models. *J. Hydrol.* **2021**, *598*, 126370. [CrossRef]
101. Ouedraogo, I.; Defourny, P.; Vanclooster, M. Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale. *Hydrogeol. J.* **2019**, *27*, 1081–1098. [CrossRef]

102. Asadollah, S.B.H.S.; Sharafati, A.; Motta, D.; Yaseen, Z.M. River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. *J. Environ. Chem. Eng.* **2021**, *9*, 104599. [CrossRef]
103. Neitsch, S.L.; Arnold, J.G.; Kiniry, J.R.; Williams, J.R. *Soil and Water Assessment Tool Theoretical Documentation Version 2009*; Technical Report; Texas Water Resources Institute: College Station, TX, USA, 2011.
104. Jung, C.; Ahn, S.; Sheng, Z.; Ayana, E.K.; Srinivasan, R.; Yeganantham, D. Evaluate River Water Salinity in a Semi-Arid Agricultural Watershed by Coupling Ensemble Machine Learning Technique with SWAT Model. *JAWRA J. Am. Water Resour. Assoc.* **2021**. [CrossRef]
105. Heddam, S.; Ptak, M.; Zhu, S. Modelling of daily lake surface water temperature from air temperature: Extremely randomized trees (ERT) versus Air2Water, MARS, M5Tree, RF and MLPNN. *J. Hydrol.* **2020**, *588*, 125130. [CrossRef]
106. Toffolon, M.; Piccolroaz, S. A hybrid model for river water temperature as a function of air temperature and discharge. *Environ. Res. Lett.* **2015**, *10*, 114011. [CrossRef]
107. Arora, B.; Şengör, S.; Spycher, N.; Steefel, C.I. A reactive transport benchmark on heavy metal cycling in lake sediments. *Comput. Geosci.* **2015**, *19*, 613–633. [CrossRef]
108. Şengör, S.S.; Spycher, N.F.; Ginn, T.R.; Sani, R.K.; Peyton, B. Biogeochemical reactive–diffusive transport of heavy metals in Lake Coeur d'Alene sediments. *Appl. Geochem.* **2007**, *22*, 2569–2594. [CrossRef]
109. Boyle, K.; Örmeci, B. Microplastics and Nanoplastics in the Freshwater and Terrestrial Environment: A Review. *Water* **2020**, *12*, 2633. [CrossRef]
110. Mihai, F.C.; Gündoğdu, S.; Khan, F.R.; Olivelli, A.; Markley, L.A.; van Emmerik, T. Chapter 11—Plastic pollution in marine and freshwater environments: Abundance, sources, and mitigation. In *Emerging Contaminants in the Environment*; Sarma, H., Dominguez, D.C., Lee, W.Y., Eds.; Elsevier: Amsterdam, The Netherlands, 2022; pp. 241–274. [CrossRef]
111. Sharma, V.K.; Ma, X.; Guo, B.; Zhang, K. Environmental factors-mediated behavior of microplastics and nanoplastics in water: A review. *Chemosphere* **2021**, *271*, 129597. [CrossRef] [PubMed]
112. Arnell, N.W.; Lowe, J.A.; Bernie, D.; Nicholls, R.J.; Brown, S.; Challinor, A.J.; Osborn, T.J. The global and regional impacts of climate change under representative concentration pathway forcings and shared socioeconomic pathway socioeconomic scenarios. *Environ. Res. Lett.* **2019**, *14*, 084046. [CrossRef]
113. Hosseiny, H.; Nazari, F.; Smith, V.; Nataraj, C. A framework for modeling flood depth using a hybrid of hydraulics and machine learning. *Sci. Rep.* **2020**, *10*, 8222. [CrossRef]
114. Nelson, J.M. iRIS Software: FaSTMECH Solver Manual. USGS, 1–36. 2013. Available online: https://i-ric.org/en/solvers/fastmech/ (accessed on 6 January 2022).
115. Andrews, F. *Hydromad Tutorial*; The Australian National University: Canberra, ACT, Australia, 2010.
116. Schoppa, L.; Disse, M.; Bachmair, S. Evaluating the performance of random forest for large-scale flood discharge simulation. *J. Hydrol.* **2020**, *590*, 125531. [CrossRef]
117. Janizadeh, S.; Pal, S.C.; Saha, A.; Chowdhuri, I.; Ahmadi, K.; Mirzaei, S.; Mosavi, A.H.; Tiefenbacher, J.P. Mapping the spatial and temporal variability of flood hazard affected by climate and land-use changes in the future. *J. Environ. Manag.* **2021**, *298*, 113551. [CrossRef]
118. Saber, M.; Boulmaiz, T.; Guermoui, M.; Abdrado, K.I.; Kantoush, S.A.; Sumi, T.; Boutaghane, H.; Nohara, D.; Mabrouk, E. Examining LightGBM and CatBoost models for wadi flash flood susceptibility prediction. *Geocarto Int.* **2021**, 1–26. [CrossRef]
119. Band, S.S.; Janizadeh, S.; Chandra Pal, S.; Saha, A.; Chakrabortty, R.; Melesse, A.M.; Mosavi, A. Flash Flood Susceptibility Modeling Using New Approaches of Hybrid and Ensemble Tree-Based Machine Learning Algorithms. *Remote Sens.* **2020**, *12*, 3568. [CrossRef]
120. Wang, Z.; Lai, C.; Chen, X.; Yang, B.; Zhao, S.; Bai, X. Flood hazard risk assessment model based on random forest. *J. Hydrol.* **2015**, *527*, 1130–1141. [CrossRef]
121. Chen, J.; Huang, G.; Chen, W. Towards better flood risk management: Assessing flood risk and investigating the potential mechanism based on machine learning models. *J. Environ. Manag.* **2021**, *293*, 112810. [CrossRef]
122. Ma, M.; Zhao, G.; He, B.; Li, Q.; Dong, H.; Wang, S.; Wang, Z. XGBoost-based method for flash flood risk assessment. *J. Hydrol.* **2021**, *598*, 126382. [CrossRef]
123. Nkiaka, E.; Taylor, A.; Dougill, A.J.; Antwi-Agyei, P.; Fournier, N.; Bosire, E.N.; Konte, O.; Lawal, K.A.; Mutai, B.; Mwangi, E. Identifying user needs for weather and climate services to enhance resilience to climate shocks in sub-Saharan Africa. *Environ. Res. Lett.* **2019**, *14*, 123003. [CrossRef]
124. Rhee, J.; Park, K.; Lee, S.; Jang, S.; Yoon, S. Detecting hydrological droughts in ungauged areas from remotely sensed hydro-meteorological variables using rule-based models. *Nat. Hazards* **2020**, *103*, 2961–2988. [CrossRef]
125. Zhang, R.; Chen, Z.Y.; Xu, L.J.; Ou, C.Q. Meteorological drought forecasting based on a statistical model with machine learning techniques in Shaanxi province, China. *Sci. Total Environ.* **2019**, *665*, 338–346. [CrossRef]
126. Hauswirth, S.M.; Bierkens, M.F.; Beijk, V.; Wanders, N. The potential of data driven approaches for quantifying hydrological extremes. *Adv. Water Resour.* **2021**, *155*, 104017. [CrossRef]
127. Manzanas, R.; Gutiérrez, J.; Fernández, J.; van Meijgaard, E.; Calmanti, S.; Magariño, M.; Cofiño, A.; Herrera, S. Dynamical and statistical downscaling of seasonal temperature forecasts in Europe: Added value for user applications. *Clim. Serv.* **2018**, *9*, 44–56. [CrossRef]
128. Li, T.; Jiang, Z.; Treut, H.L.; Li, L.; Zhao, L.; Ge, L. Machine learning to optimize climate projection over China with multi-model ensemble simulations. *Environ. Res. Lett.* **2021**, *16*, 094028. [CrossRef]

129. Ayzel, G. Machine Learning Reveals a Significant Shift in Water Regime Types Due to Projected Climate Change. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 660. [CrossRef]
130. Perrin, C.; Michel, C.; Andréassian, V. Improvement of a parsimonious model for streamflow simulation. *J. Hydrol.* **2003**, *279*, 275–289. [CrossRef]
131. Abatzoglou, J.T.; Brown, T.J. A comparison of statistical downscaling methods suited for wildfire applications. *Int. J. Climatol.* **2012**, *32*, 772–780. [CrossRef]
132. Fisher, A.; Rudin, C.; Dominici, F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *J. Mach. Learn. Res.* **2019**, *20*, 1–81.
133. Trenberth, K.E. Climate change caused by human activities is happening and it already has major consequences. *J. Energy Nat. Resour. Law* **2018**, *36*, 463–481. [CrossRef]
134. Naumann, G.; Alfieri, L.; Wyser, K.; Mentaschi, L.; Betts, R.A.; Carrao, H.; Spinoni, J.; Vogt, J.; Feyen, L. Global Changes in Drought Conditions Under Different Levels of Warming. *Geophys. Res. Lett.* **2018**, *45*, 3285–3296. [CrossRef]
135. Seibert, J.; Strobl, B.; Etter, S.; Hummer, P.; van Meerveld, H.J.I. Virtual Staff Gauges for Crowd-Based Stream Level Observations. *Front. Earth Sci.* **2019**, *7*, 70. [CrossRef]
136. Fienen, M.N.; Lowry, C.S. Social.Water—A crowdsourcing tool for environmental data acquisition. *Comput. Geosci.* **2012**, *49*, 164–169. [CrossRef]
137. Wu, D.; Del Rosario, E.A.; Lowry, C. Exploring the Use of Decision Tree Methodology in Hydrology Using Crowdsourced Data. *JAWRA J. Am. Water Resour. Assoc.* **2021**, *57*, 256–266. [CrossRef]