hydrology

IMPACT FACTOR 3.1

CITESCORE 4.9

Review

Protocols for Water and Environmental Modeling Using Machine Learning in California

Minxue He, Prabhjot Sandhu, Peyman Namadi, Erik Reyes, Kamyar Guivetchi and Francis Chung

Special Issue

Advancing Hydrological Science Through Artificial Intelligence: Innovations and Applications

Edited by

Dr. Xiaofeng Liu, Dr. Yi Hong, Dr. Ryan C. Johnson, Dr. Md Shahabul Alam and Dr. Quan Guo

MDPI

https://doi.org/10.3390/hydrology12030059

hydrology

MDPI

Review

Protocols for Water and Environmental Modeling Using Machine Learning in California

Minxue He *, Prabhjot Sandhu, Peyman Namadi, Erik Reyes, Kamyar Guivetchi and Francis Chung

California Department of Water Resources, 1516 9th Street, Sacramento, CA 95814, USA
* Correspondence: kevin.he@water.ca.gov

Abstract: The recent surge in popularity of generative artificial intelligence (GenAI) tools like ChatGPT has reignited global interest in AI, a technology with a well-established history spanning several decades. The California Department of Water Resources (DWR) has been at the forefront of this field, leveraging Artificial Neural Networks (ANNs), a core technique in machine learning (ML), which is a subfield of AI, for water and environmental modeling (WEM) since the early 1990s. While protocols for WEM exist in California, they were designed primarily for traditional statistical or process-based models that rely on predefined equations and physical principles. In contrast, ML models learn patterns from data and require different development methodologies, which existing protocols do not address. This study, drawing on DWR's extensive experience in ML, addresses this gap by developing standardized protocols for the development and implementation of ML models in WEM in California. The proposed protocols cover four key phases of ML development and implementation: (1) problem definition, ensuring clear objectives and contextual understanding; (2) data preparation, emphasizing standardized collection, quality control, and accessibility; (3) model development, advocating for a progression from simple models to hybrid and ensemble approaches while integrating domain knowledge for improved accuracy; and (4) model deployment, highlighting documentation, training, and open-source practices to enhance transparency and collaboration. A case study is provided to demonstrate the practical application of these protocols step by step. Once implemented, these protocols can help achieve standardization, quality assurance, interoperability, and transparency in water and environmental modeling using machine learning in California.

Keywords: machine learning; protocols; water and environmental modeling; California

Academic Editors: Xiaofeng Liu, Yi Hong, Ryan C. Johnson, Md Shahabul Alam and Quan Guo

Received: 17 January 2025
Revised: 8 March 2025
Accepted: 12 March 2025
Published: 14 March 2025

Citation: He, M.; Sandhu, P.; Namadi, P.; Reyes, E.; Guivetchi, K.; Chung, F. Protocols for Water and Environmental Modeling Using Machine Learning in California. *Hydrology* **2025**, *12*, 59. https://doi.org/10.3390/hydrology12030059

1. Introduction

Water and environmental systems comprise surface water bodies like rivers and lakes, along with groundwater aquifers, wetlands, and other components. Currently, these systems confront pressing challenges. These challenges include but are not limited to intensifying weather and hydrological extremes, habitat degradation, climate change, and unsustainable use of natural resources. California, in particular, is highly susceptible to these challenges due to its intricate and aging water infrastructure, along with a large population and economy that place significant strain on its resources and infrastructure [1,2]. Addressing these challenges requires a comprehensive understanding of complex environmental processes and their interactions, which is where water and environmental modeling plays a pivotal role.

Water and environmental modeling offers a multifaceted approach to comprehending complex systems by integrating diverse data sources, representing complex processes,

and facilitating simulation and prediction. It also serves as a valuable tool for informing decision making across various sectors by providing insights into the current and projected conditions of the systems and their responses to different scenarios. Moreover, it facilitates communication and collaboration among stakeholders by providing a common platform for analyzing data, evaluating options, and fostering productive dialogue, ultimately contributing to more informed decision making [3,4].

California has a rich history of developing and employing models to plan and manage the State's intricate water and environmental systems. For example, as far back as 1930, the California Division of Water Resources (now known as the California Department of Water Resources (DWR), the overarching mission of which is to sustainably manage the water resources of California in cooperation with other agencies) issued Bulletin 120, which provides forecasts for April to July runoff volume based on various variables, such as precipitation, snowfall, and runoff indices [5,6]. In 1931, the DWR devised the Bulletin 27 Salinity Intrusion Model to estimate the required flows for managing seawater intrusion in the Sacramento–San Joaquin Delta (Delta), the hub of California's water transfer systems [7]. In 1978, the DWR introduced DAYFLOW, a straightforward mass balance model intended for determining the daily historical hydrology along the boundaries of the Delta [8]. Subsequently, in 1986, the DWR developed DWRSIM, a water planning model, to simulate the joint operation of the State Water Project (SWP) and Central Valley Project (CVP) [9]. Later on, the DWR and the U.S. Bureau of Reclamation collaborated to develop CalSim as a replacement for DWRSIM [10]. In the early 1990s, the DWR implemented the hydrodynamics and water quality model DSM2, designed to simulate variables such as flow, stage, salinity, water temperature, and other water quality variables in the Delta [11].

These models, encompassing statistical and physics-based approaches, play a vital role in various aspects of water management. They facilitate real-time flood and water supply forecasting [12–17], drought assessment and management [18,19], regulatory compliance [20], water rights management [21], climate change adaptation and mitigation planning [22,23], infrastructure investment and design [24], ecosystem restoration [25–27], and long-term water resource planning [28], among others. Given the interconnected nature of water systems, ecosystems, and human activities, computer simulations are indispensable to comprehensively understanding and predicting their interactions. Consequently, the majority of significant water projects in California rely on analyses conducted through modeling [9,10,28].

As models become more central to managing California's water systems, scrutiny and controversy surrounding their use and development are also rising. It is widely understood among the State's water community that the adoption of standardized modeling protocols is imperative. These protocols can help enhance the quality of models and modeling studies, while also bolstering the trust and confidence of decision makers and stakeholders who rely on model outcomes to inform their decisions. In light of this, the Bay-Delta Modeling Forum (now California Water and Environmental Modeling Forum (CWEMF)) developed a landmark document in 2000 outlining protocols for water and environmental modeling in California [3]. The CWEMF updated this document in 2021 to reflect advancements in the field, driven by new questions, technologies, and increased stakeholder engagement [4]. The updated document provides current best practices and highlights emerging techniques like machine learning, which could significantly enhance modeling capabilities. However, it does not offer specific guidance on developing and applying machine learning models. Machine learning techniques have advanced significantly in recent years, leading to a surge in their applications across California's water and environmental sectors. These applications span surface water hydrology [29–44], groundwater hydrology [45–52], hydrometeorology/hydro-climatology [53–64], water quality [65–85], ecology [86–98], and water

management and operations [99–109], among others. Machine learning models offer a distinct advantage over traditional statistical methods by automatically discovering intricate patterns and interactions within data, without demanding predefined assumptions (e.g., linearity and normality) about the functional forms of relationships [110,111]. Furthermore, while physics-based models necessitate a thorough understanding of underlying physical mechanisms and precise parameterization, machine learning can extract valuable insights even when a complete, theoretical understanding of the system is lacking [112]. These capabilities render machine learning particularly effective in analyzing complex environmental systems, where variable interactions are frequently non-linear, interdependent, and challenging to express through conventional methods. By employing data-driven learning, machine learning can reveal obscured patterns, enhance prediction accuracy, and augment existing modeling approaches, especially in scenarios where traditional techniques encounter limitations. As a result, the protocols governing traditional models and machine learning models may not always be interchangeable.

This paper aims to supplement existing modeling principles and guidelines [3,4,113,114] by providing protocols for the development and implementation of ML models in water and environmental modeling in California. These protocols stem from practical insights gained from using conventional machine learning techniques in water and environmental modeling. Although generative artificial intelligence (GenAI) [115] represents the latest machine learning advancement, its application in California's water modeling community remains limited. As such, the guidelines provided in this paper are not specifically tailored for GenAI models. More specifically, the protocols proposed in this study focus on conventional machine learning models that process quantitative and spatial water-related data, including numerical or categorical meteorological, hydrological, water quality, biological, and geospatial data. These protocols are not applicable to machine learning models designed for processing text, image, audio, or video data.

The remainder of the paper is structured as follows: Section 2 reviews machine learning in water and environmental modeling, covering the typical life cycle of a machine learning model, the roles that machine learning can play in water and environmental modeling, and the DWR's three decades of experience with machine learning. Section 3 outlines the machine learning protocols based on the DWR's experience and includes a case study to demonstrate these protocols step by step. Section 4 discusses the benefits and caveats of implementing these protocols. Section 5 concludes the paper and suggests future directions.

2. Machine Learning in Water and Environmental Modeling

Artificial intelligence (AI), a branch of computer science that emerged in the 1950s, aims to develop machines capable of simulating human intelligence [116]. Over time, its limitations spurred the growth of machine learning (ML), a subfield of AI that emphasizes statistical modeling and data-driven pattern recognition. ML has evolved significantly, gaining momentum with the advancement of neural networks since their inception [117], followed by the rise of deep learning (DL) and, more recently, generative AI (GenAI), fueled by increasing computational power and data availability. An overview of AI and ML history is provided in Appendix A. In addition, the terminologies used in this paper are defined in Appendix B. This section presents the lifecycle of typical ML models, their role in water and environmental modeling, and a recap of our three decades' experience in ML model development and application in water and environmental modeling in California.

2.1. The Life Cycle of a Machine Learning Model

The life cycle of a typical ML model encompasses several key stages, including problem definition, data preparation, model development, and mode deployment (Figure 1). Data preparation can be divided into data collection and data pre-processing. Model development can be further divided into model selection, model training, and model evaluation. During the problem definition phase, the specific problem or task that the model aims to address is clearly defined, along with the objectives and desired outcomes. Following problem definition, the process moves to data collection, where relevant datasets are gathered from various sources, ensuring they are comprehensive and representative of the problem domain. Subsequently, data pre-processing is carried out to clean, transform, and prepare the data for analysis, addressing issues such as missing values, outliers, and noise. It is important to note that feature engineering, which involves extracting and transforming features from raw data for use in ML models, is typically considered part of the ML model life cycle [118]. However, for simplicity, this paper broadly categorizes it under data pre-processing.

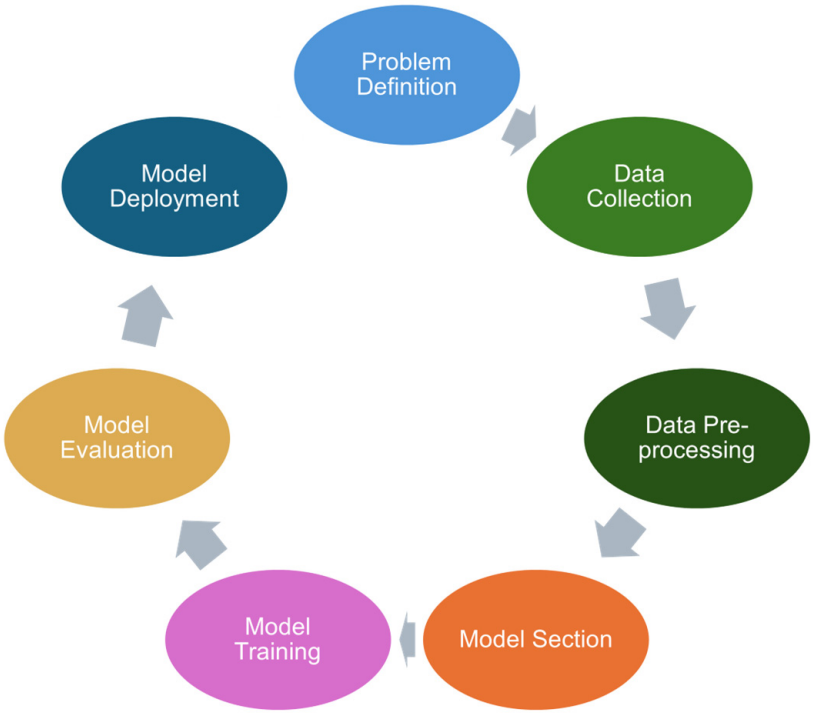

Figure 1. The life cycle of a typical machine learning model.

Once the data are pre-processed, the next step is model selection, where the most appropriate ML algorithms are chosen based on the problem type, data characteristics, and desired outcomes. With the data pre-processed and the models selected, the models are trained and tested, through which they learn patterns and relationships between features (model input) and the label which is the target variable or the output that an ML model is trained to predict, referring to the definitions in Appendix B. Specifically, datasets are typically divided into training, validation, and test sets to ensure that a model generalizes well to unseen data. The training set is used to learn patterns, the validation set helps tune hyperparameters (configurable variables set before training that influence the model's learning process but are not learned from the data) and prevent overfitting, and the test set evaluates final model performance. When datasets are small, cross-validation—where data are repeatedly split into training and validation sets—offers a more reliable estimate of model performance by reducing dependence on a single split. A key challenge in this

process is overfitting, where a model learns noise or specific patterns in the training data rather than general trends, leading to poor performance on new data. Out-of-sample testing, such as evaluation on a separate test set, helps detect overfitting by revealing discrepancies between training and real-world performance. To mitigate overfitting, regularization techniques such as L1/L2 penalties [119,120], dropout in neural networks, pruning in Decision Trees, and data augmentation are commonly used to ensure the model captures meaningful patterns rather than memorizing the training data. Model performance can be evaluated via statistical metrics and/or visual inspection. Based on the evaluation results, adjustments may be made to the model or its parameters to optimize performance further. Finally, upon satisfactory evaluation outcomes, the model is deployed into production, making it available for real-world use, where it can provide insights, make predictions, or automate decision-making processes, thus completing the life cycle of an ML model. Throughout this cycle, iteration and continuous improvement are key, as models may require updating or retraining to maintain relevance and effectiveness over time. Further elaboration on each of these steps will be provided in later sections.

2.2. Roles of Machine Learning Models in Water and Environmental Modeling

The complexities of water and environmental systems along with a changing climate pose challenges for traditional modeling approaches. ML models have emerged as powerful tools, offering unique capabilities that complement and enhance existing methods (e.g., [121–124]). A number of roles that ML models play in water and environmental modeling are described as follows.

(1) Enhanced simulation and prediction

ML models excel at identifying patterns and relationships within datasets, often improving prediction accuracy compared with traditional methods, even with a small-to-moderate number of data. However, DL models typically require very large datasets to achieve high performance. These capabilities enable ML to support accurate simulations and predictions of future environmental conditions. This includes tasks like the following:

(a) Hydrological forecasting: Forecasting precipitation (e.g., [55,56,58,125,126]), reservoir inflows (e.g., [40,127]) and outflows [43], water level (e.g., [128,129]), flood events and flood risks (e.g., [130–134]), flood inundation area (e.g., [135]), seasonal water supply forecast (WSF) [34–36,136], and drought events (e.g., [137]) with improved accuracy and lead time.

One prime example is the multi-model machine learning metasystem (M^4), developed for the U.S. Department of Agriculture's Natural Resources Conservation Service (NRCS) [34]. M^4 enhances the traditional principal components regression approach used by the NRCS by incorporating supervised and unsupervised machine learning techniques, ensemble learning, and evolutionary optimization. A follow-up study [35] validated M^4 by testing it across 20 hindcasting cases at 11 locations with diverse hydro-climatic conditions, including the Truckee River at Farad, California. The results demonstrated improved WSF accuracy, enhanced geophysical explainability, and strong operational feasibility, supporting M^4's adoption for large-scale water resource management applications.

Building upon the above work, Fleming et al. [36] investigated the integration of next-generation satellite remote sensing-based snow data into WSF models in coordination with the U.S National Aeronautics and Space Administration's Jet Propulsion Laboratory in Pasadena, California. Specifically, the study assessed the spatially and temporally complete (STC) MODSCAG fractional snow-covered area (fSCA) product as a complement to traditional in situ snow data. Testing was conducted in four diverse watersheds, including one in California, using both conventional NRCS models and the M^4 framework. Results indicated that incorporating fSCA data improved forecast accuracy, particularly for short-lead,

late-season predictions when snowlines retreat above ground-based measurement sites. In addition, the ML-based M4 model is better equipped to leverage satellite data compared with traditional linear statistical models. The study further highlighted the operational advantages of integrating satellite-derived snow data into M^4, with significant implications for enhancing water resource management in snow-dependent regions, including key headwater watersheds in California.

(b) Water quality modeling: Forecasting the spatial and temporal variations in water quality variables [138–140], including salinity (e.g., [66,67]), water temperature (e.g., [65,85,141]), ion concentrations (e.g., [75,76]), sediment (e.g., [73]), algal bloom [142,143], and dissolved oxygen [66], among others; identifying potential contamination sources; and assessing the effectiveness of mitigation strategies. To provide an example, Rath et al. [82] enhanced an empirical salinity model for the San Francisco Bay–Delta estuary by integrating a Bayesian ANN and new inputs. The hybrid model improved performance and characterized prediction uncertainty. This work demonstrated a practical method for integrating existing models with Bayesian ANNs, constrained by system behavior insights, for broad environmental modeling applications.

(c) Climate change analysis: Understanding the complex dynamics of climate systems and predicting future trends in hydro-climatic patterns (e.g., [144–148]). For instance, Wi et al. [42] evaluated the physical realism of ML models for projecting streamflow under climate change. By comparing a long short-term memory (LSTM) model to process-based models across California watersheds, the study found that while LSTM can capture hydrologic responses, these models may produce unrealistic runoff projections in warming scenarios. Incorporating process model outputs as additional inputs improves realism, but effectiveness depends on the process model used. The findings revealed both the potential and limitations of ML in hydrologic forecasting under a changing climate.

(2) Uncovering hidden insights

ML algorithms can analyze a vast number of water and environmental data, including satellite imagery, sensor readings, and historical records. This ability to process complex data allows them to uncover hidden patterns and relationships that might be missed or even misrepresented by traditional methods. These insights can be crucial to the following:

(a) Identifying emerging environmental threats: Early forecasting and detection of flood or drought (e.g., snow drought) events and changes in water quality (e.g., [149–151]). To illustrate, Tanim et al. [107] developed an ML-based flood detection framework that integrated satellite imagery and change detection methods to identify urban flooding in San Diego, California. By combining Sentinel-1 satellite data with supervised and unsupervised classification techniques, the study effectively mapped flooded areas with high accuracy. The use of unsupervised methods enabled rapid flood identification with minimal data requirements. This approach enhanced early forecasting and detection capabilities, enabling city planners and emergency responders to mitigate flood risks and improve urban resilience to extreme weather events.

(b) Developing targeted conservation strategies: Identifying crucial habitats (e.g., [152]), predicting species (e.g., fish, birds, etc.) distribution (e.g., [153]), and informing conservation efforts (e.g., [154,155]). For example, Olivetti et al. [92] integrated computational fluid dynamics and ML to analyze how migratory Chinook salmon navigates complex flow environments. By decomposing movement into passive drift and active locomotion, the study improved predictions of migration trajectories and energy expenditure. The findings offered valuable insights for conservation strategies, enabling better habitat management, informed water flow regulation, and improved infrastructure design to support migratory species under changing environmental conditions.

(c) Optimizing resource planning and management: Understanding the factors influencing water availability and demand, system vulnerability, leading to more sustainable water use practices and water resource planning (e.g., [106,156]). As an example, Robinson et al. [106] developed an ML-based early warning system to detect long-term water supply vulnerabilities under climate change. Using historical and projected hydro-climatic data from Northern California's reservoir system, the study applied classification models to predict future supply risks at lead times of up to 20 years. Results indicated that ML classifiers outperformed the benchmarks, with predictive accuracy declining at longer lead times. The findings underscored the potential for data-driven early warning signals to inform adaptive water resource planning.

(3) Efficiency and scalability

ML models offer significant advantages in terms of efficiency and scalability. They can process large datasets rapidly, enabling model users to conduct simulations and explore various scenarios in a shorter timeframe. This allows for the following:

(a) Real-time operations and decision making: Providing near-real-time insights into water and environmental conditions, facilitating proactive responses to emerging issues (e.g., [157,158]). By way of example, Tillotson et al. [96] developed an ML-based forecasting tool to reduce the risk of entrainment loss for endangered salmonids at large-scale water diversions in California's Sacramento–San Joaquin Delta. By using a quantile regression forest approach trained on historical data, the model predicted weekly entrainment risk. The study also developed an interactive web-based application that allowed for real-time operational adjustments, enhancing decision making for water management while balancing ecological conservation with water supply needs.

(b) Simulating large-scale water and environmental processes: Modeling complex interactions between different components of the environment, such as the impact of climate change on water resources or the spatial and temporal variations in critical water quantity or water quality variables (e.g., [159,160]). For instance, Kratzert et al. [38] applied ML, specifically LSTM networks, to improve regional rainfall–runoff modeling across 531 basins, including 29 in California. By leveraging large-sample datasets, the LSTM models effectively captured hydrologic variability and outperformed traditional models in streamflow prediction. The study advanced large-scale water and environmental modeling by demonstrating how ML can integrate meteorological, hydrological, and physical basin characteristics to improve regional hydrologic simulations.

(4) Integration with traditional models

ML models are not necessarily intended to replace traditional models used in water and environmental modeling. Instead, they can serve as a complementary approach or, in some cases, provide an alternative when they offer improved performance. They can also be integrated with existing models for the following aims:

(a) Improve the accuracy, speed, and efficiency of simulations (e.g., [70,80,83]). For example, Qi et al. [80] developed an ML-based dashboard to mimic the DSM2 model for salinity simulation at critical locations in the Delta. This dashboard operates notably quicker than DSM2, and it is capable of running on a mobile phone with internet connectivity. The outcomes can be swiftly generated and displayed within seconds.

(b) Provide a deeper understanding of the complex interactions within environmental systems (e.g., [161–164]). For instance, Ransom et al. [81] developed a hybrid ML approach to predict and visualize nitrate contamination in the Central Valley aquifer, California, by integrating traditional physically based models with Boosted Regression Tree. By incorporating outputs from existing models, the ML approach provided deeper insights into the factors influencing nitrate concentrations in groundwater. The study highlighted the

significance of redox conditions, nitrogen flux, and groundwater age in shaping contamination patterns. This integration of ML with traditional modeling frameworks improved predictive accuracy and supported more effective groundwater management strategies.

2.3. Machine Learning Experience of the California Department of Water Resources

The California Department of Water Resources (DWR) has been developing and applying ML tools for water and environmental modeling since the early 1990s. Initial efforts primarily focused on salinity modeling [165], as the salinity level is a critical factor influencing water operations in the Delta, the State's central water hub. Recent work has expanded to include the modeling of specific ion constituents, including dissolved chloride, bromide, and sulfate [75]. Additionally, ML models have been applied to simulate other water quality variables, including water temperature [166], suspended sediment [73], trihalomethane [167], and flow-related parameters such as carriage water [168] and streamflow [169]. These applications are illustrated in Figure 2 and summarized as follows.

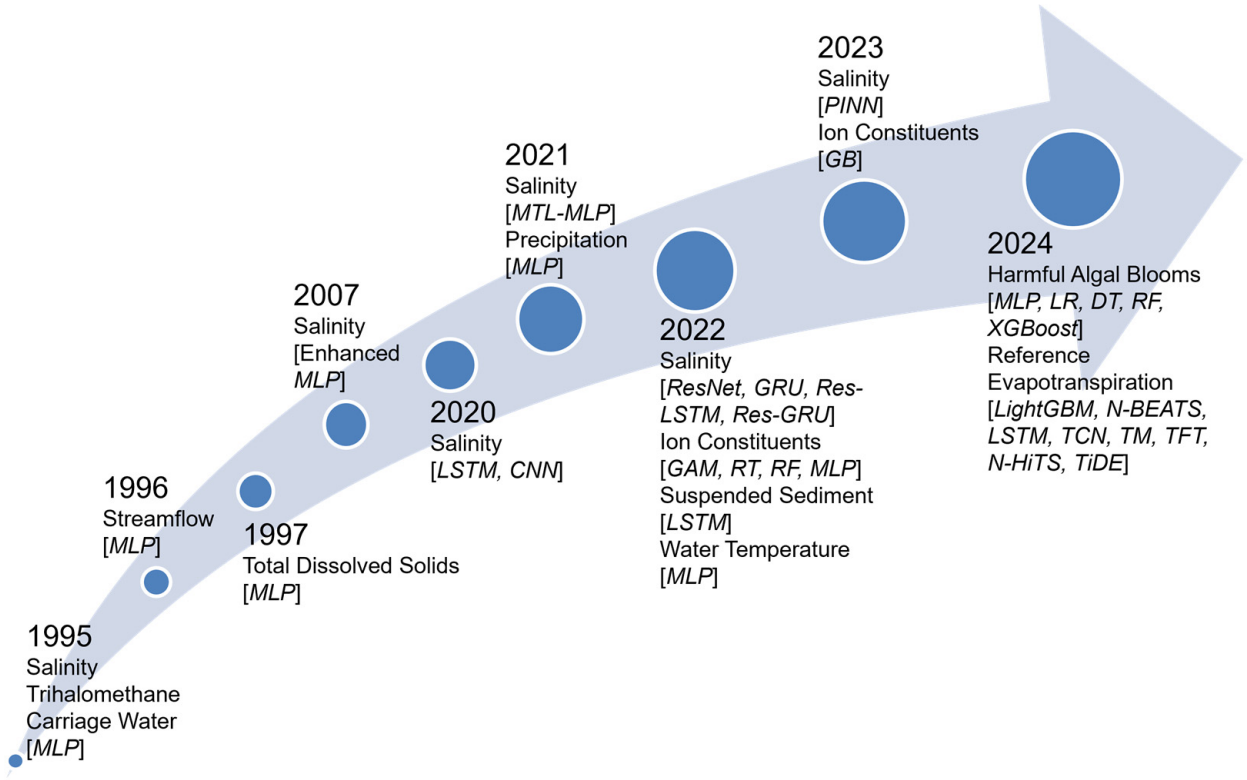

Figure 2. Target variables modeled by using machine learning (ML) and the corresponding ML models developed to simulate them in the DWR. MLP: Multi-Layer Perceptron; LSTM: long short-term memory; CNN: Convolutional Neural Network; MTL-MLP: multi-task learning MLP; ResNet: Residual Neural Network; GRU: Gated Recurrent Unit; Res-LSTM: Residual LSTM; Res-GRU: Residual GRU; GAM: Generative Additive Model; RT: Regression Tree; RF: Random Forest; PINN: Physics-Informed Neural Network; GB: Gradient Boosting; LR: Logistic Regression; XGBoost: Extreme Gradient Boosting; LightGBM: Light Gradient-Boosting Machine; N-BEATs: Neural Basis Expansion Analysis for Interpretable Time Series Forecasting; TCN: Temporal Convolutional Network; TM: Transformer Model; TFT: Temporal Fusion Transformer; N-HiTs: Neural Hierarchical Interpolation for Time Series Forecasting; TiDE: Time-Series Dense Encoder.

2.3.1. Applications in Modeling Salinity

The Delta plays a vital role in California's water supply system, necessitating effective management of ocean-driven salinity intrusion to ensure water quality for municipal, agricultural, and environmental uses. Early applications of ML for salinity modeling in

the Delta involved the development of separate Multi-Layer Perceptron (MLP) models—a type of Artificial Neural Network (ANN)—for each salinity monitoring location [96]. These models, with one or two hidden layers, demonstrated superior performance compared with traditional empirical methods. Subsequent enhancements focused on (a) improving model robustness by training them in diverse hydrological and operational scenarios [170], (b) optimizing input feature selection and training strategies [170–172], and (c) simplifying integration into operational planning tools [71,170,173,174].

A major advancement came with the adoption of multi-task learning (MTL), allowing a single model to predict salinity at multiple locations simultaneously [77–79]. The advent of deep learning further revolutionized salinity modeling with architectures such as Convolutional Neural Networks (CNNs) and Long-Short-Term Memory (LSTM) networks [70]. CNNs effectively captured peak salinity events, while LSTM excelled at managing long-term dependencies in time-series data [70]. Further refinements led to innovative architectures, including Residual Networks (ResNets), Res-Gated Recurrent Units (Res-GRUs), and Res-LSTM models [78,79]. To overcome data limitations and improve generalizability, data augmentation was employed to generate diverse datasets representing a wider range of hydro-climatic and operational conditions [80]. Transfer learning was also implemented, enabling models trained on simulated datasets to predict real-world salinity observations [80].

These advancements were complemented by the development of an interactive dashboard, designed to support decision making [80]. This dashboard offers an intuitive interface for water managers to visualize salinity levels and assess compliance in various planning scenarios, reflecting different operational strategies or hydro-climatic conditions.

2.3.2. Applications in Modeling Ion Constituents

Ion constituents are critical indicators of water quality, determining its suitability for various uses. Traditionally, regression models estimate ion constituent concentrations based on salinity, represented by electrical conductivity (EC). However, these models often struggle with ion constituents exhibiting strong non-linear relationships with EC. To address this, the DWR developed ML models as alternatives.

The first study [75] focused on simulating 12 ion constituents in the South Delta: total dissolved solids, dissolved chloride, dissolved sulfate, dissolved sodium, dissolved calcium, dissolved magnesium, dissolved nitrate, dissolved potassium, dissolved bromide, dissolved boron, alkalinity, and hardness. By using data from seven stations, the study compared different ML models, including Regression Tree (RT), Random Forest (RF), and ANNs. Random Forest generally outperformed the other models in the study.

Building on this foundation, the second study [76] expanded the scope to 30 stations across the Delta, incorporating additional predictors beyond EC, such as the Sacramento River X2 position, month of the year, and water year type, to improve prediction accuracy. This study evaluated RT, RF, Gradient Boosting, and ANNs, finding that ANNs performed the best overall. Additionally, an interactive, web-based dashboard powered by the ANN model was developed and hosted on Microsoft Azure. This dashboard allows users to modify input parameters and visualize resulting ion concentration predictions, providing a valuable tool for data-driven water quality management in the Delta. This effort will be discussed in further detail in Section 3.

2.3.3. Applications in Modeling Other Variables

In addition to salinity and ion constituents, the DWR has developed ML models to simulate minimum Delta outflow (MDO) and marginal export cost (MEC) [168,169,175–177]. MDO represents the freshwater flow required to maintain salinity standards at specific

compliance locations in the Delta, particularly during increased water exports. MEC quantifies the additional flow needed to offset salinity increases caused by a unit of export. Traditional methods often oversimplified Delta hydrodynamics, struggling to capture the complex interplay of flows, gate operations, and salinity. The DWR's ANN models address this by leveraging multiple inputs, including river inflows, export pumping rates, Delta Cross-Channel gate positions, and tidal influences, to predict salinity more accurately. These ANN-based approaches have been integrated into operational planning tools like DWRSIM and CALSIM [169,173,174], offering a faster and more precise alternative to traditional methods for simulating MDO and MEC.

ML models have also been developed to emulate traditional log-linear models for simulating trihalomethane (THM) formation in the Delta [167]. These models use inputs such as bromide concentration, dissolved organic carbon (DOC), UV-254, chlorine dose, ammonia, reaction time, temperature, and pH. Validation demonstrated their superior accuracy compared with traditional approaches.

Moreover, the DWR developed LSTM models to estimate suspended sediment concentration (SSC) at key locations in the Delta [73]. The LSTM models demonstrated better or comparable performance compared to the traditional process-based DSM2 model, particularly in capturing temporal variability and peak magnitudes in SSC. The models utilized flow, stage, and SSC data for training and testing, with additional analysis incorporating precipitation and wind speed. The study highlighted the potential for ML to complement existing process-based methods, offering faster and more flexible solutions for sediment management in complex estuarine systems like the Delta.

Additionally, ML models have been developed to estimate water temperature at boundary locations in the Delta using air temperature and flow as input variables [166]. In particular, MLPs with a single hidden layer were employed for this purpose. These models demonstrated a high level of accuracy in predicting water temperatures, effectively capturing the relationships between the input parameters and the observed outcomes.

Furthermore, a set of DL models were developed to forecast monthly reference evapotranspiration at 55 weather stations in California, some of which are located in the Delta. The work was built upon the studies [53,54] but extended beyond them in terms of (a) exploring the improvement of DL models in global versus local learning; and (b) automating hyperparameter optimization to achieve the most desirable forecasting accuracy.

ML models have also been applied to assess the risks of harmful algal blooms (HABs) in the Delta [178]. Specifically, five models were developed to simulate HAB risk levels (low or high) based on nine environmental variables across 16 locations. Among these models, the Random Forest model demonstrated the best performance, achieving the highest accuracy, precision, and recall.

The protocols presented in this paper are primarily based on our three decades of experience in this field, as mentioned above. However, we acknowledge that numerous other machine learning research efforts and applications in water and environmental modeling have been conducted in California outside of the DWR. Since the goal of this paper is not to provide an exhaustive review of all ML studies in the State, these studies are instead compiled and presented in Appendix C.

3. Machine Learning Protocols and Case Study

Drawing on the DWR's extensive experience outlined in Section 2.3, this section provides an overview of the key principles and guidelines for the four main steps of ML model development and application: problem definition, data preparation, model development, and model deployment. A case study on using ML to model ion constituents is also included to illustrate these steps in practice.

3.1. Problem Definition

3.1.1. Problem Definition Protocols

Problem definition is the cornerstone of the successful application of ML in all fields, including water and environmental modeling. It acts as a compass, guiding the entire process and ensuring efficient resource allocation, optimal model selection, and ultimately, the development of appropriate tools for addressing critical water and environmental challenges. The typical protocols to craft a clear and well-defined problem statement involves the following:

(1) Understand the Context
   (a) Identify the problem: Begin by clearly identifying the specific water-related or environmental issues to be addressed. Is it water data anomaly detection, water quality deterioration, flood, drought, and water supply forecasting, or others?
   (b) Stakeholder needs: Consider the needs and interests of stakeholders who might benefit from the model's insights, such as policymakers, water managers, or the public.

(2) Define the Scope
   (a) Specificity: Avoid broad and ambiguous statements like "improve water quality understanding". Instead, focus on specific and measurable objectives.
   (b) Temporal and spatial considerations: Specify the desired prediction timeframe (e.g., hourly, daily, seasonal, or annual) and the spatial scope (e.g., a specific river, channel, an entire watershed, or groundwater basin).

(3) Define the Desirable Outcome
   (a) Measurable goals: State what you want the model to achieve in a quantifiable manner. This could involve predicting specific water quality parameters, identifying areas of water quality degradation, or forecasting the probability of specific events.
   (b) Performance metrics: Define the metrics you will use to evaluate the model's success in achieving the desired outcome. This could involve metrics including but not limited to accuracy, precision, and recall for classification tasks or bias, Mean Absolute Error, and the coefficient of determination for regression tasks.

(4) Refine and Iterate
   (a) Seek feedback: Share the draft problem statement with colleagues or experts in the field and stakeholders to gather feedback and ensure clarity and feasibility.
   (b) Iterative process: Be prepared to make adjustments based on feedback and emerging information throughout the project.

3.1.2. Problem Definition Example

The composition of water in the Delta is significantly influenced by various salts and minerals, such as chloride, sulfate, sodium, magnesium, and potassium. Collectively identified as ions, these salts and minerals are pivotal to maintaining the ecological balance and agricultural viability and ensuring the safety of drinking water in the Delta.

The traditional approach to measuring these ion constituents involves infrequent (e.g., roughly monthly) on-site sampling (i.e., grab samples) at limited locations followed by laboratory analysis. This method, while accurate, is time-consuming and expensive and provides results that are not immediately available and lack spatial comprehensiveness. The necessity for an advanced approach is underscored by the challenges inherent in sampling methods, prompting the exploration of more efficient alternatives. Measurements

of salinity (represented by EC) show great promise because they correlate with ions and are easily accessible from automatic sensors installed at key locations throughout the Delta. Historically, the DWR has utilized parametric regression equations for converting EC into ion levels [179]. However, the performance of these conversion equations is limited for ions having strong non-linear relationships with EC.

The objective of this case study is two-fold. Firstly, it seeks to create ML models that can mimic and possibly enhance the current regression equations for EC–ion conversion in terms of both goodness-of-fit metrics and visual inspection. Secondly, it provides a user-friendly interactive dashboard based on these ML models. Through its interactive features, the dashboard allows stakeholders, water managers, and the public to easily modify input parameters and quickly visualize ion concentrations for various input scenarios, without needing any programming skills.

3.2. Data Preparation

3.2.1. Data Preparation Protocols

Effective data preparation, including data collection and pre-processing, is indispensable to advancing ML model development in water and environmental modeling. By preparing diverse, high-quality datasets and adhering to best practices for data preparation, developers can build robust ML models that provide valuable insights into water and environmental processes and thus inform decision making. To ensure the quality and reliability of the collected data, it is essential to adhere to best practices throughout the data preparation process. The following outlines some typical practices:

(1) Use of standardized protocols: It is necessary to adhere to standardized protocols and methodologies for data collection, measurement, model simulation, and sampling to ensure consistency and comparability across datasets.
(2) Implementation of quality assurance/quality control (QA/QC) measures: It is necessary to implement rigorous QA/QC procedures to identify and rectify errors, outliers, and inconsistencies in the collected data, including the calibration of instruments, duplicate measurements, and cross-validation with reference data.
(3) Selection of representative data: It is necessary to choose sampling locations that are representative of the spatial and temporal variability of the water and environmental processes under study, taking into account factors such as study purpose, study period, and data availability, among others.
(4) Data accessibility and documentation: It is necessary to document metadata, data sources, and collection methods to facilitate data sharing, reproducibility, and transparency.
(5) Data pre-processing: This process involves examining raw input data and transforming them—either directly or through feature extraction—into a form that ML models can effectively utilize. Typical techniques for pre-processing include normalization and standardization. Normalization scales the data to a range, typically [0, 1], by adjusting values proportionally based on the minimum and maximum of each feature. This is particularly useful when features have varying scales. Standardization transforms data to have a mean of 0 and a standard deviation of 1, making them suitable for ML algorithms that assume normally distributed features.
(6) Data split: Once the data are prepared, they are typically divided into separate subsets. A common approach involves a three-way split into training, validation, and test sets, where the validation set is used for tuning hyperparameters and preventing overfitting. In some cases, cross-validation is employed instead of a fixed validation set to ensure robust model evaluation. These strategies help optimize the model's parameters and assess its performance on unseen data.

3.2.2. Data Preparation Example

This case study combines data from three sources, spanning measurements from 1959 to 2022 across 30 monitoring locations in the Delta. Figure 3 shows these monitoring locations, which are grouped into three sub-regions based on their distinct hydrological and environmental characteristics: the Old-Middle River (OMR), the San Joaquin River Corridor, and the South Delta. This regional classification captures the varying water sources and quality impacts throughout the Delta.

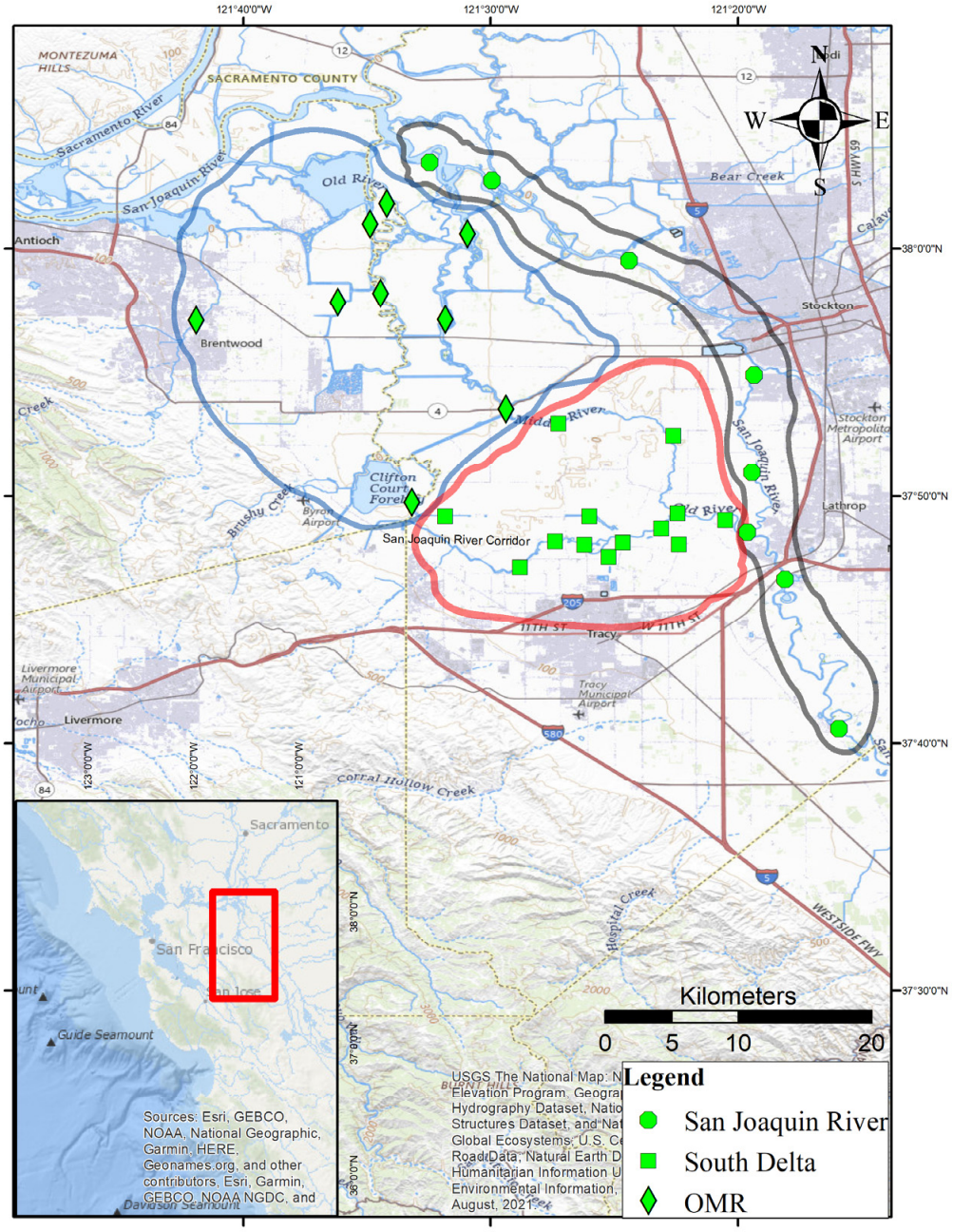

Figure 3. Study ion sample locations in the Delta.

The primary dataset originates from historical records utilized by [180] which have been quality-controlled and well documented, containing ion grab samples, EC, and X2 position data collected between 1959 and 2018 at 19 stations within the study area. The second dataset includes samples collected by the DWR between 2018 and 2020, focusing on seven stations in the South Delta. Finally, the third dataset, downloaded from the

California Data Exchange Center (CDEC) website (https://cdec.water.ca.gov/, accessed on 1 March 2024), extends the geographical and temporal coverage by including samples from 13 stations within the interior Delta, collected from 2018 to 2022.

The integration of these datasets provides a comprehensive view of the water quality dynamics within the Delta, capturing a broad spectrum of hydrologic conditions and environmental changes over time. This approach ensures a robust analytical foundation, enabling the development of models to accurately simulate ion levels under a variety of conditions.

Table 1 presents a detailed overview of the ion constituents analyzed in this study, highlighting the sample size, range of data, and the standard deviation (SD) of the data for each ion constituent. This table illustrates the dataset's composition and variability, offering insights into the distribution and concentration of ions across the Delta. The nine ions presented in the table each play a unique role in influencing the water quality and ecological health of the Delta.

Table 1. Overview of study ion samples *.

Ion	Sample Size	Data Range	SD	Period	Units
TDSs	1466	49–2120	204	1968–2022	mg/L
Mg^{2+}	1336	2–102	8.6	1959–2022	mg/L
Na^{+}	1575	6–343	44	1959–2022	mg/L
Ca^{2+}	1335	5.8–244	18	1959–2022	mg/L
Cl^{-}	1972	4–775	77	1959–2022	mg/L
$SO_4{}^{2-}$	1066	5–350	46.5	1959–2022	mg/L
Br^{-}	1239	0.01–2.3	0.22	1990–2022	mg/L
Alkalinity	1036	26–198	27.6	1959–2020	mg/L as $CaCO_3$
K^{+}	1148	0.87–11	1.35	1959–2022	mg/L

* Adapted from [76].

The data pre-processing stage is crucial to preparing the dataset for ML analysis, ensuring that the input data are clean, normalized, and suitable for the ML algorithms to process efficiently. The following outlines the key pre-processing steps undertaken, including visualization, normalization, and the treatment of categorical data.

Visualization plays a pivotal role in understanding the underlying patterns and relationships within the dataset. Figure 4 illustrates the relationships between EC and the ion constituents across the three sub-regions of the Delta. These visual representations help identify the linear, bifurcated, and non-linear relationships between EC and ion levels, facilitating the categorization of ions into three distinct groups. This indicates that the relationship between EC and ion concentrations can vary significantly depending on the ion type and environmental factors. The groups are defined as follows:

Group 1: Ions with a linear relationship to EC, indicating a direct and proportional increase in ion concentration with EC. These ions comprise magnesium (Mg^{2+}) and total dissolved solids (TDSs)).

Group 2: Ions exhibiting bifurcation or branching patterns in response to EC, suggesting varying ion concentration pathways under different conditions. These ions feature sodium (Na^{+}), chloride (Cl^{-}), and calcium (Ca^{2+}).

Group 3: Ions with non-linear relationships to EC, reflecting complex dynamics that cannot be captured by a straightforward linear model. These ions consist of sulfate ($SO_4{}^{2-}$), alkalinity, potassium (K^{+}), and bromide (Br^{-}).

The division into these groups allows for a more nuanced analysis, acknowledging the diverse behaviors of ion constituents in the Delta's waters and ensuring that the ML models are optimally configured to reflect these complexities.

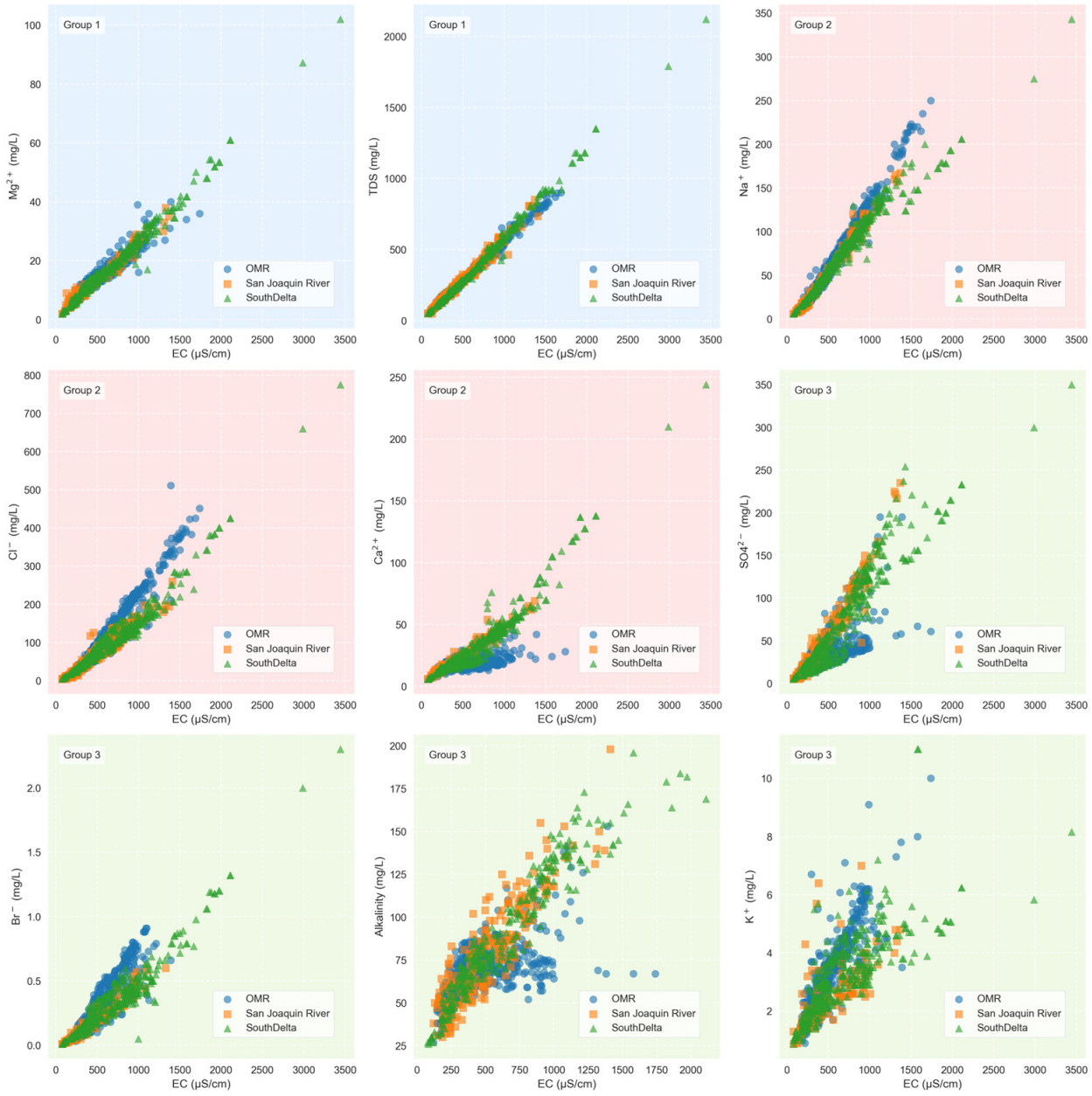

Figure 4. Scatter plots of salinity (represented by EC on *x*-axis) and study ion constituents (*y*-axis). Ion constituents are categorized into three groups according to their relationships with salinity.

To accommodate the diverse range observed across the dataset, particularly in the numerical predictors of EC and X2 position, a normalization process is helpful. This normalization adjusts the dataset to a uniform scale, specifically between 0 and 1, for these predictors. The procedure puts all features on a comparable scale, crucial to preventing biases from differences in measurement units or ranges. High-value ions, or any disproportionately large measurements, could otherwise skew the ML models' performance.

The incorporation of categorical variables, such as Water Year Type (WYT), month, and sub-region, requires appropriate encoding techniques to ensure efficient use by ML algorithms. A commonly used method is one-hot encoding [181], which represents each category as a unique binary vector. However, depending on the specific modeling requirements and dataset characteristics, alternative techniques, such as multi-hot encoding [182], may also be employed to capture additional categorical relationships more effectively. For instance, WYT, denoting the hydrological condition of a year with classifications ranging

from "wet" to "critical", required conversion into a computationally understandable format. To achieve this, one-hot encoding was utilized, transforming these categorical variables into binary columns [33,38]. With five water year types, we created four binary columns, where the fifth category (critical) is implicitly represented when all other columns are zero. Each column represents a potential category value, converting qualitative attributes into a quantitative framework suitable for analysis. For example, a "'wet" year is encoded as [1, 0, 0, 0], a "dry" year as [0, 0, 0, 1], and a "critical" year as [0, 0, 0, 0]. This method allows for the seamless integration of essential hydrological, temporal, and spatial data into the models without compromising the numerical requirements of the ML algorithms.

3.3. Model Development

3.3.1. Model Development Protocols

Selecting the appropriate ML architecture is crucial to developing effective models in water and environmental modeling. Based on the DWR's previous ML model development experience, we outline several best practices for ML architecture selection, ranging from simple models to ensemble and hybrid methods, tailored to the unique challenges and requirements of environmental applications.

(1) Start Simple

It is recommended to begin with simple ML models, more interpretable models, such as basic versions of Multi-Layer Perceptron (MLP) models [183] or Decision Trees [184], to establish a baseline understanding of the data and the problem at hand. These models can help identify key relationships between input variables and the target variable while providing insights into feature importance.

(2) Ensemble Methods

Ensemble methods can be considered to improve model performance and robustness. These techniques include Random Forest [185] and Gradient Boosting [186], which combine multiple base models to create a more generalized and accurate learner. Additionally, techniques like bootstrap aggregation (bagging) can be applied to MLPs and other models to enhance stability and reduce overfitting.

(3) Hybrid Methods

Hybrid methods that integrate multiple ML architectures or combine ML with traditional modeling approaches should be explored [83,187–189]. Hybrid methods leverage the strengths of different techniques to address specific challenges in water and environmental modeling. For example, combining physics-based models with data-driven ML models can improve predictive accuracy while retaining interpretability and physical realism.

(4) Consider Spatial and Temporal Dynamics

The spatial and temporal dynamics inherent in environmental data should be accounted for when selecting ML architectures. Convolutional Neural Networks (CNNs) [190] are well suited for spatial data analysis, capturing spatial patterns and relationships. Recurrent Neural Networks (RNNs) and long short-term memory (LSTM) networks [191] excel at modeling temporal dependencies, making them suitable for time-series data in water and environmental modeling.

(5) Incorporate Domain Knowledge

Domain knowledge and expert insights should be integrated into ML architecture selection to ensure that the models capture relevant environmental processes accurately. Domain-specific constraints, physical laws, and expert understanding can guide the choice of architectures and enhance the relevance and applicability of the models.

Training and evaluating ML models in water and environmental modeling necessitates meticulous attention to multiple factors to guarantee their resilience, dependability, and precision. Drawing from the DWR's prior experience, we deem the subsequent practices indispensable:

(1) Sensitivity Analysis

Sensitivity analysis is a critical step in water and environmental ML modeling to identify the influence of individual input variables on model predictions. By systematically varying input parameters, the features that have the most significant impact on outcomes can be identified. This process helps prioritize data collection efforts, improve model robustness, and ensure interpretability in applications such as streamflow prediction, water quality assessment, and ecosystem modeling. Sensitivity analysis also aids in detecting potential overfitting to less important features, fostering more reliable and generalizable ML models.

(2) Overfitting Prevention

Overfitting occurs when an ML model learns patterns specific to the training data, including noise or irrelevant details, leading to poor generalization to new, unseen data. Regularization methods, such as L1 and L2 regularization, dropout, pruning, data augmentation, bagging, and early stopping [119,120,192], help prevent overfitting by penalizing overly complex models. In water and environmental modeling, where datasets may be limited or noisy, regularization techniques ensure that models generalize well to unseen data. Regularization should be implemented to balance model complexity and performance, enhancing the model's reliability and interpretability.

(3) Ensemble Learning

Ensemble learning techniques combine multiple models to improve predictive performance and robustness. In water and environmental modeling, where uncertainties and complexities are prevalent, ensemble methods offer enhanced predictive capabilities. Ensemble techniques like bagging [192], boosting [193], or stacking [194] are utilized to leverage the diversity of individual models and achieve more accurate predictions across different environmental scenarios.

An emerging approach is the use of higher-level ensembles, where structurally diverse models are combined to improve predictive accuracy and robustness. One example is the previously discussed M^4 metasystem [34–36]. Another example is the ensemble modeling framework introduced by Umirbekov et al. [195] for seasonal streamflow forecasting. Their approach integrates generalized linear models, Gaussian processes, support vector regression (SVR), and Random Forest. The predictions from these diverse models are then aggregated using an additional SVR model as a metalearner, enabling the ensemble to leverage the strengths of each modeling approach. By increasing structural variability among ensemble members, this method enhances stability and reduces the risk of overfitting to any single modeling paradigm.

(4) Cross-Validation

Cross-validation is a valuable technique for assessing model performance and generalizability, especially when datasets are too limited to support traditional training–validation–testing splits. This is particularly relevant in water and environmental modeling, where data may be limited and often exhibit temporal or spatial dependencies. Methods such as k-fold cross-validation or stratified cross-validation [196] can be used to partition the data, providing an alternative form of out-of-sample testing. By leveraging cross-validation, ML models can be evaluated more robustly, helping to mitigate overfitting and provide reliable performance estimates across diverse environmental conditions.

(5) Hyperparameter Optimization

Hyperparameter optimization [197] involves systematically tuning model hyperparameters to maximize performance while avoiding overfitting or underfitting. In water and environmental modeling, where model complexity may vary based on the intricacies of environmental processes, hyperparameter optimization is crucial to fine-tuning model performance. Techniques such as grid search, randomized search, gradient-based optimization, Bayesian optimization, evolution strategies, and hyperband [197–205] can be utilized to identify the optimal hyperparameter settings for ML models.

These best practices aid in streamlining the development and application of ML models for water and environmental modeling, ensuring their consistency, interoperability, and applicability in addressing the State's water and environmental challenges.

3.3.2. Model Development Example

(1) Model Selection

The primary objective of this case study was to predict ion concentrations in the Delta, a problem that inherently falls within the domain of regression analysis in ML. Regression problems require models that can predict continuous outcomes, such as ion concentrations, based on one or more predictor variables, such as EC and X2 position. Given this context, the model selection process was guided by the necessity to choose ML models adept at solving regression problems and capable of handling the complexity and non-linearity of environmental data.

Moreover, it is important to note that our dataset, derived from grab samples, does not constitute a time series. Grab samples provide a snapshot of water quality at specific points in time and locations but lack the sequential, time-dependent structure that characterizes time-series data. As a result, models that are good at time-series forecasting, such as LSTM networks, are not the most appropriate models for our dataset. LSTM and similar time-series models excel at capturing temporal dependencies and patterns over time, which our grab sample dataset does not present.

Given these considerations, the following ML models were selected for their proven effectiveness in regression problems and their suitability for the dataset's characteristics.

Decision Trees: Decision Trees offer a straightforward, interpretable approach to regression, making them a useful starting point for understanding the relationships within our data [184,206]. They allow us to visualize how decisions are made based on the input features, although they may not capture complex relationships as effectively as more sophisticated models.

Random Forest (RF): As an ensemble of Decision Trees, Random Forest mitigate some of the overfitting issues associated with individual trees and are more capable of handling complex, non-linear relationships in regression problems [185]. Their robustness and ability to deal with high-dimensional data make them well suited for predicting ion concentrations.

Gradient-Boosting Machines (GBMs): GBMs build models sequentially to correct the residuals of prior models, gradually improving prediction accuracy [193]. This technique is highly effective for regression, especially when dealing with non-linearity and interactions between predictors, as is common in environmental datasets.

Artificial Neural Networks (ANN): The Artificial Neural Network (ANN) is capable of capturing complex and non-linear relationships within datasets [183]. At its core, a typical ANN consists of interconnected units called neurons, which are organized into layers: an input layer, several hidden layers, and an output layer. Neurons in these layers interact through connections that are weighted, with the network's learning process revolving around adjusting these weights based on the input data.

Hidden layers are the ANN's workhorse, allowing the model to learn features at different levels of abstraction. The activation function within each neuron determines how signals are transformed as they pass through, introducing non-linearity into the model and enabling it to learn complex patterns. Common activation functions include ReLU, sigmoid, and tanh, each bringing different properties to the model's learning capability. Additionally, the learning rate is a crucial parameter that governs the step size during the weight update process, influencing the model's convergence speed and accuracy.

In this study, the ANN architecture incorporates four hidden layers. This setup aims to leverage the ANN's general abilities to process and analyze the data, utilizing neurons, activation functions, and an optimized learning rate to enhance model performance. A comparative summary of the ML models used in this study is presented in Table A2 in Appendix D.

(2) Model Training

In this case study, the input–output datasets are randomly divided, allocating 80% for training purposes and the remaining 20% for evaluation. This partitioning facilitates the assessment of each ML model's predictive accuracy and generalization capability, providing a clear indication of performance on unseen data.

To prevent overfitting—a scenario where a model learns the training data too well, compromising its performance on new data—an early-stopping function is incorporated into the training process, with a threshold set to 50 epochs. Early stopping acts as a form of regularization, effectively halting the training when no improvement in the model's performance on a validation set is observed for a consecutive number of epochs specified by the threshold. This technique ensures that the model maintains its ability to generalize from the training data to unseen data, thus safeguarding against overfitting.

The evaluation and comparison of the ML models are grounded in their performance metrics on the evaluation dataset. The models are assessed based on their R-squared (R^2) and Mean Absolute Error (MAE), with a close examination of the convergence between their performance on the training dataset to further guard against overfitting. Such scrutiny ensures that the selected models not only exhibit strong predictive accuracy but also demonstrate robust generalization capabilities across different data samples.

This structured approach to data division, model evaluation, and the strategic incorporation of early stopping underscores the meticulousness of the study's methodology. It aims to ensure the development of reliable and generalizable ML models for predicting ion concentrations in the Delta.

Among the ML models evaluated, the ANN model demonstrated superior performance in predicting ion concentrations, particularly for ions exhibiting non-linear relationships with EC. This finding led to the selection of the ANN model for comparison against classical methods.

Once the model is selected, the next step is the optimization of the ANN hyperparameters, which is a computationally demanding task, particularly when these ANN models incorporate multiple hidden layers. This study's ANN models, which include four hidden layers, required a robust hyperparameter search to fine-tune each model's settings for predicting various ion concentrations accurately.

Hyperparameter search is a pivotal step in ML that involves experimenting with different combinations of model settings. These settings, or hyperparameters, such as the number of hidden layers, neurons in each layer, activation functions, and learning rate, are crucial to defining the model's capacity to learn from data. The optimal hyperparameters are those that contribute to the lowest evaluation error, indicating the model's effectiveness in making accurate predictions.

Given the complexity of searching hyperparameters for ANN models with four hidden layers, the computational workload exceeds the capabilities of standard desktop computers. To address this challenge, this study leveraged Microsoft Azure cloud computing resources, utilizing an environment equipped with 96 CPU cores. This powerful computing infrastructure significantly accelerated the hyperparameter search process, allowing for more extensive and rapid exploration of the hyperparameter space. This expeditious approach was instrumental in identifying the configurations that most effectively minimized the prediction error.

The detailed outcomes of the hyperparameter search, particularly the optimal number of neurons and activation functions chosen for each ion constituent model, are presented in Table A3 in Appendix D.

(3) Model Evaluation

(a) K-fold Cross-validation

The K-fold cross-validation analysis was conducted to assess and evaluate the generalization capability of the developed ANN models. Unlike a simple train–evaluation split, K-fold cross-validation systematically uses different portions of the data for training and evaluation. This approach helps mitigate the risk of overfitting and ensures that the model's performance is not dependent on a particular division of the data. It is particularly valuable in scenarios where the available dataset is not extensive, maximizing the use of data for both training and evaluation.

In this case study, the dataset was divided into five equally sized subsets, or folds (K = 5). The model training and evaluation process was then repeated five times, with each fold used exactly once as the evaluation set, while the remaining four folds were used for training. This method allowed every data point in the dataset to be used for both training and evaluation, providing a thorough evaluation of the model's predictive accuracy.

By using a fixed random seed for reproducibility, the subsets were created randomly to ensure that each fold was a representative sample of the whole dataset. This randomized division is crucial to minimizing bias and variance in the model evaluation process.

The K-fold cross-validation indicates consistent accuracy of the ANN models in predicting ion concentrations, as evidenced by the evaluation metric MAE calculated for each fold in comparison to the corresponding MAE values of the benchmark model (Figure A2 in Appendix D).

(b) Sensitivity Analysis

Sensitivity analysis is a critical component in model validation, particularly for systems as complex and variable as ion constituent levels in the Delta. Figures 5 and 6 showcase the model's predictions for bromide (Br) and chloride (Cl) concentrations, respectively, across three sub-regions, i.e., Old-Middle River (OMR), San Joaquin River Corridor, and South Delta, under specified conditions (April, Wet Year, and Sacramento X2 = 75 km). These figures highlight the models' ability to adjust simulations in response to changing levels of EC.

Both figures depict a smooth gradient in the simulated ion concentrations as EC values vary, reflecting the models' nuanced response to input changes. This is indicative of an effective balance between sensitivity and stability; the models are responsive to EC changes without displaying erratic behavior for minor fluctuations. Such stability in the presence of varied input conditions is essential to obtaining reliable predictions in environmental contexts, where data can naturally exhibit variability.

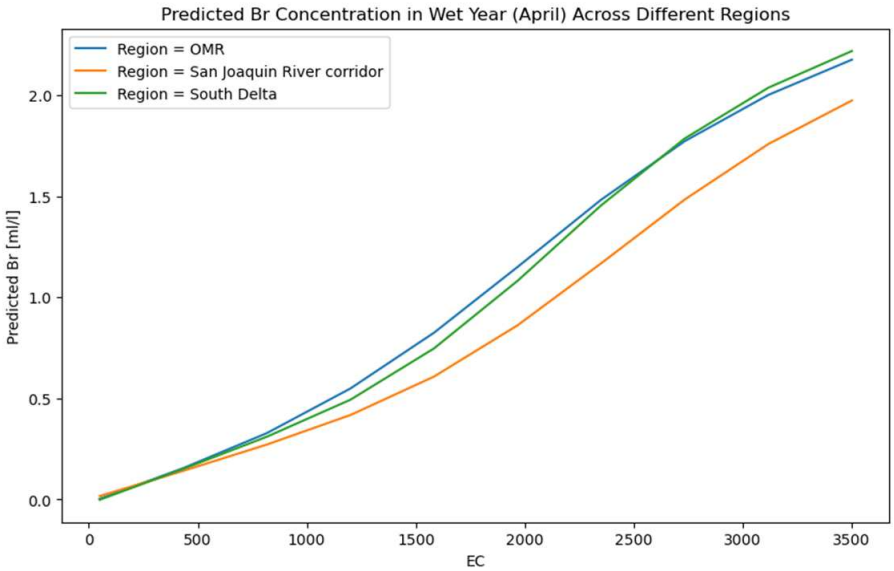

Figure 5. Sensitivity of bromide concentration to EC in the Delta: a Wet Year scenario in April.

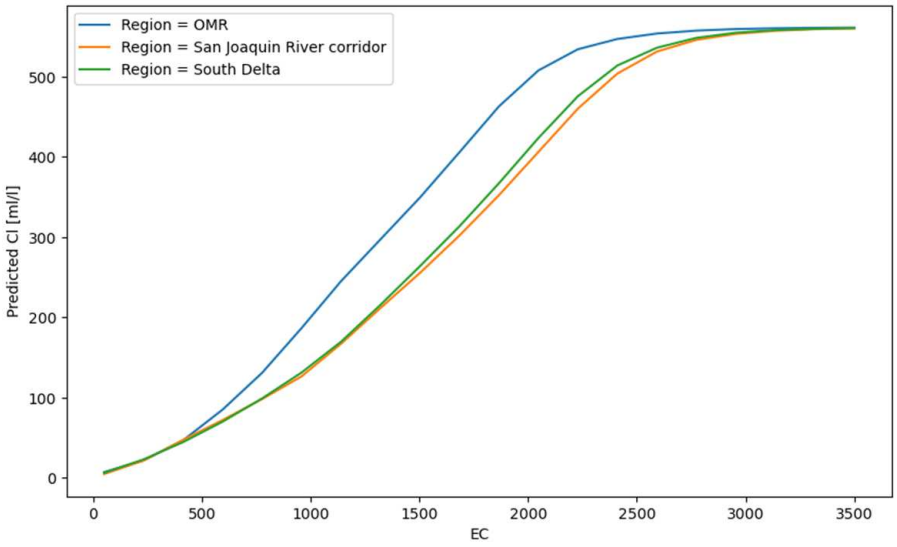

Figure 6. Sensitivity of chloride concentration to EC in the Delta: a Wet Year scenario in April.

The sensitivity analysis as demonstrated in these figures informs the practical application of these ANN models. For ML model users, the consistency and smoothness in predictions signify that the models can serve as dependable tools for ion simulation. These characteristics ensure that the outputs are accurate and interpretable, enabling informed decision making. The models' generalizability, indicated by their performance under specific test conditions, is crucial to applications across the diverse ecological landscape of the Delta.

(c) Performance Evaluation

The performance of the ANN models is quantitatively superior to that of traditional parametric regression models, particularly for Group 3 ion constituents, which exhibit more complex and non-linear relationships with EC (Figure 7). The improvements are evident in both R^2 values and MAE. For Group 1 ions, the improvement in R^2 is limited due to the strong linear relationship between EC and the ion constituents, although the improvement in MAE is notable. For both Group 2 and Group 3 ions, improvements in both metrics are evident. The most remarkable improvements are observed for dissolved sulfate, with an 85% improvement in R^2 and a 59% improvement in MAE. This indicates that the

ANN model provides much better simulations for ions where the traditional regression model struggles to capture the strong non-linear relationships between the predictand and the predictors.

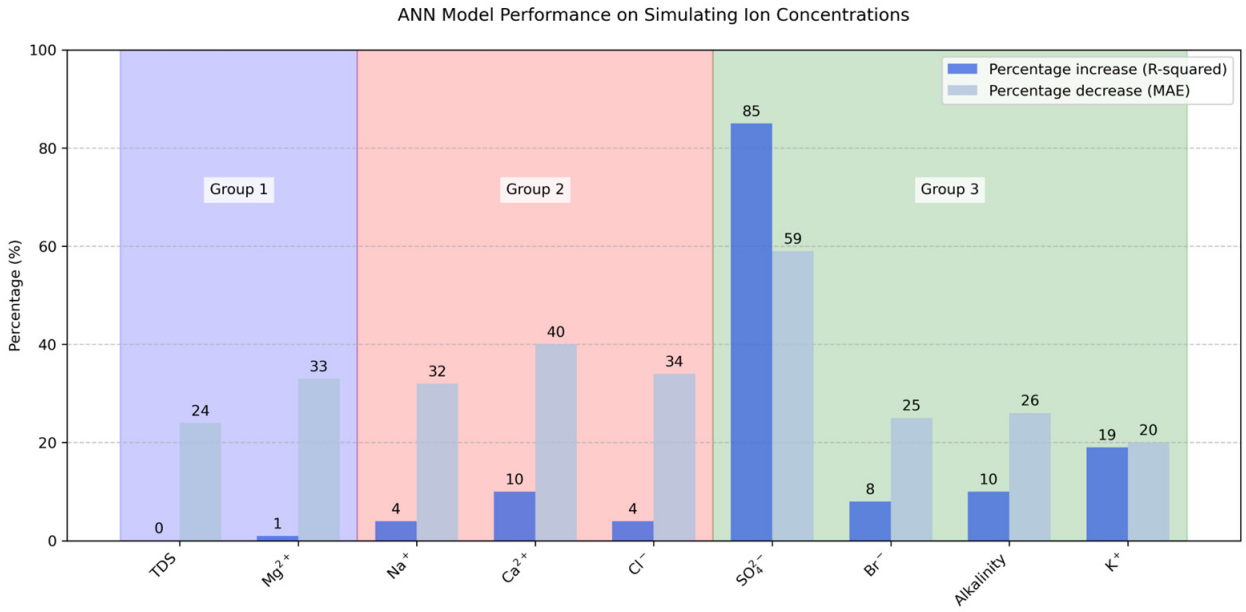

Figure 7. ANN model performance in simulating the concentrations of nine ion constituents based on percent improvement from the benchmark model represented by R^2 and MAE. Each metric represents the average value obtained across the five-fold cross-validation.

3.4. Model Deployment

3.4.1. Model Deployment Protocols

Deploying ML models effectively is imperative for their successful application in water and environmental modeling. By adhering to protocols and best practices such as comprehensive documentation, hands-on training, open-source collaboration, engagement with the user community, publication in peer-reviewed journals, and maintenance and improvement, model developers can ensure transparency, collaboration, and innovation. These practices facilitate informed decision making, enhance stakeholder engagement, and contribute to the sustainable management of water resources and environmental conservation efforts. The following delineates some of these practices we consider essential, drawn from the DWR's past experiences.

(1) Comprehensive Documentation

Thorough documentation is paramount to understanding and replicating ML model deployment. It is recommended to provide a detailed record of the model architecture, data pre-processing steps, hyperparameters, and assumptions in a technical document. Additionally, a comprehensive user manual with detailed instructions is required for deployment, usage, and troubleshooting to aid users in effectively utilizing the ML model. Transparent documentation promotes collaboration and facilitates knowledge transfer among stakeholders.

(2) Open-Source Collaboration

Open-sourcing the ML model code and associated documentation encourages collaboration and promotes transparency. Sharing code on open-source platforms like GitHub enables peer review, feedback, and contributions from the wider community. Open-source collaboration fosters innovation, accelerates model development, and ensures reproducibility in water and environmental modeling.

(3) Hands-On Training

Hands-on training sessions are essential to educating users on model deployment, interpretation, and application. These interactive sessions offer practical experience and guidance for using ML models in real-world scenarios. Customizing training materials for different user groups—such as engineers, scientists, managers, and stakeholders—ensures relevance and maximizes effectiveness.

(4) User Community Engagement

Engaging the user community can be facilitated through brown bag sessions and user group meetings focused on water and environmental modeling. These gatherings serve as platforms for knowledge sharing, collaboration, feedback collection, and networking among stakeholders.

(5) Peer-Reviewed Journal Articles

Publishing research findings and methodologies in peer-reviewed journal articles disseminates outcomes and contributes to the scientific community. Peer-reviewed publications lend credibility and validation to ML model development and application in water and environmental modeling. It also shares insights, challenges, and lessons learned to advance the field and promote transparency in model deployment and communication.

(6) Maintenance and Improvement

After deploying ML models, it is crucial to maintain and improve them. Maintenance includes updating the model with new data, retraining it when performance declines, and addressing any issues related to computational infrastructure or data pipeline failures. Improving the model requires ongoing evaluation and incorporation of user feedback, new methodologies, or additional features to enhance its accuracy and applicability. Regular audits should also be conducted to assess compliance with regulatory and ethical standards, ensuring the model remains relevant and trustworthy in addressing evolving challenges in California's water and environmental systems. This proactive approach ensures the long-term success and sustainability of ML applications.

3.4.2. Model Deployment Example

We deployed the ML models developed in the case study through two parallel approaches. First, for technical users who wish to use and customize the models, we have made the source code of the models (in Python) and data available in a public GitHub repository (https://github.com/CADWRDeltaModeling/Ion_Study_Dashboard/tree/main, accessed on 1 March 2024).

Second, for users who prefer not to interact with the models directly, we designed and implemented an interactive ion simulation dashboard tool, as illustrated in Figure 8. This tool combines pre-trained ML models (RF, RT, GB, and ANN) with established regression methods to provide ion predictions across the three zones of the interior Delta. The deployment leverages a cloud-based workflow, with ML models maintained in a GitHub codebase and executed through Microsoft Azure's computational infrastructure. The live platform is accessible via https://dwrdashion.azurewebsites.net/Dashboard (accessed on 1 March 2024). The interface prioritizes accessibility through intuitive design elements. Users interact with the platform by using dynamic controls, including sliding scales for EC values, contextual menus for site selection and water year classification, and additional parameter selectors. The "Compute" functionality generates instant visual outputs, displaying comparative predictions across all modeling approaches. This setup allows users to thoroughly explore various hydrological scenarios, examining how changes in

EC, Sacramento X2, monthly variations, and water year type impact ion concentrations in the Delta.

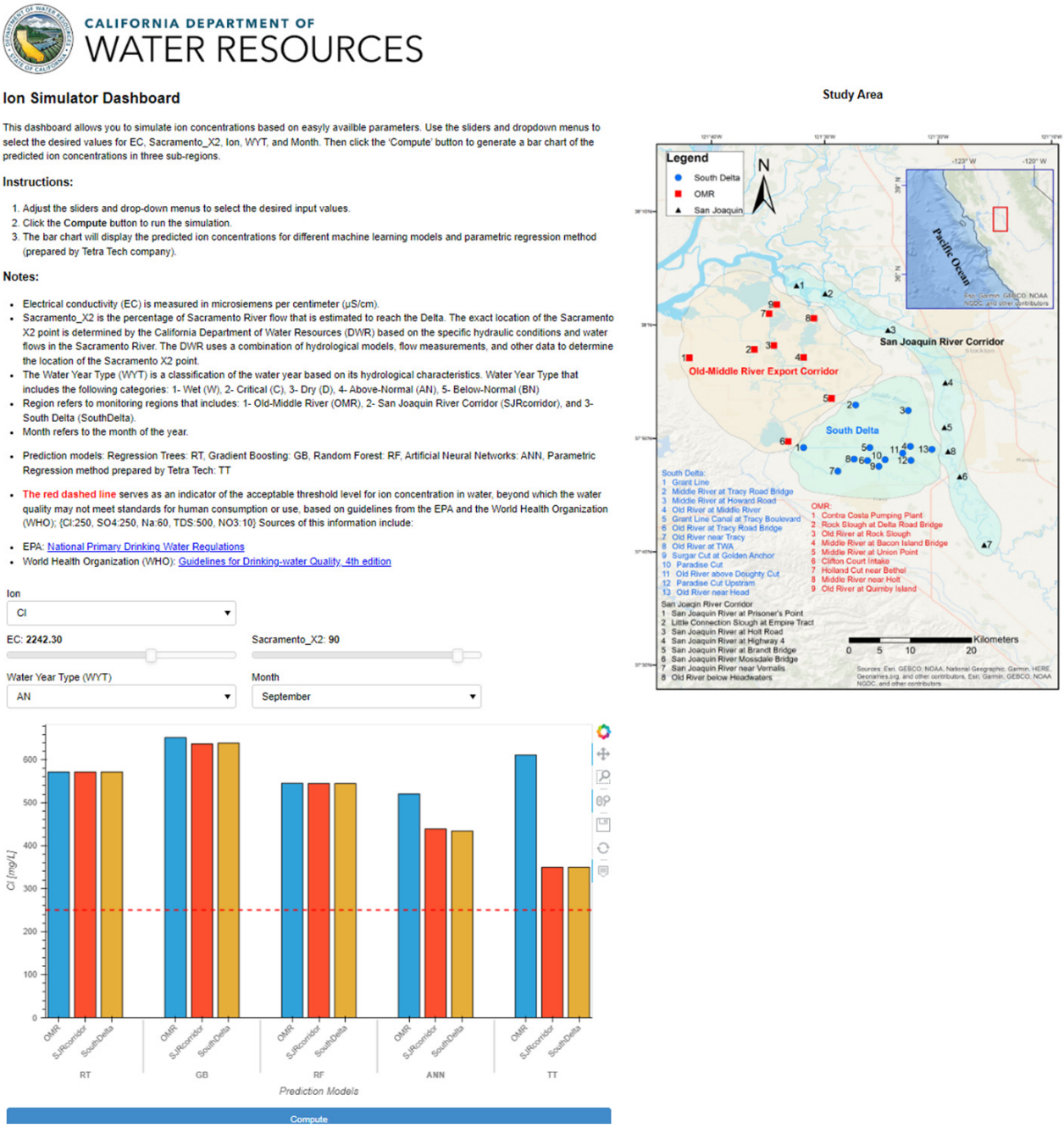

Figure 8. Screenshot of the interactive Delta ion simulator dashboard interface, displaying the results of the four ML models (ANN, RT, RF, and GB) in simulating nine ion constituents in the Delta region given user-specified inputs.

The methodology and findings from the case study have been comprehensively documented, peer-reviewed, and published [75,76], ensuring transparency and accessibility for a wider audience. Both the ML models developed and the interactive dashboard tool have been showcased at multiple forums, including stakeholder meetings, technical conferences, and user group sessions, to promote awareness and engagement within the community. Based on the feedback received, we have updated the model and dashboard to enhance their practical applicability. We will continue monitoring their performance and making improvements as new questions, data, and techniques emerge. Additionally, a hands-on

training session will be organized to guide users through effectively utilizing the dashboard tool. This session aims to provide practical experience, address user-specific questions, and demonstrate the tool's capabilities in simulating and analyzing ion predictions for water and environmental modeling.

4. Discussion

As ML garners increasing attention and finds more applications within the California water community, the ML protocols proposed in the current study become imperative for several distinct benefits.

(1) Standardization: These protocols provide standardized guidelines and procedures for ML model development and deployment. Standardization ensures consistency and reproducibility across different projects and research efforts, facilitating comparison and collaboration among water and environmental model developers, users, and stakeholders in California.

(2) Quality assurance: These protocols help ensure the quality and reliability of ML models by specifying rigorous criteria for data quality, model performance, and validation methods. By adhering to established protocols, model developers can mitigate common pitfalls such as overfitting, resulting in more robust and trustworthy modeling outcomes.

(3) Interoperability: The protocols enhance interoperability by establishing standardized guidelines on data formats, model development, evaluation, and deployment. Following these guidelines ensures that ML models can seamlessly integrate with existing modeling frameworks, databases, and decision support systems. This promotes collaboration among agencies and researchers, improving the consistency, usability, and scalability of ML-based solutions for water and environmental modeling in California.

(4) Transparency and reproducibility: These protocols promote transparency and reproducibility by documenting the entire modeling process, including data sources, preprocessing steps, model architectures, hyperparameters, and evaluation metrics. Transparent and reproducible models allow stakeholders to understand and validate modeling decisions, ensuring accountability and trustworthiness in decision-making processes.

(5) Capacity building: These protocols facilitate capacity building by providing training materials, best practices, and guidelines for model developers, users, and stakeholders involved in ML-based water and environmental modeling. A training workshop is being scheduled to daylight the protocols and provide hands-on training on how to implement them. The workshop is expected to empower stakeholders with the knowledge and skills needed to effectively apply ML techniques in real-world water and environmental applications in California.

(6) Continual improvement: Protocols support continual improvement by fostering feedback mechanisms, community engagement, and peer review processes. Iterative refinement of protocols based on feedback from stakeholders and advances in ML research ensures that protocols remain relevant, adaptable, and responsive to evolving challenges and opportunities in water and environmental modeling.

While following these protocols benefits the broader water community in California when developing and applying ML models, there are several caveats to consider. First, California's water and environmental systems are inherently complex, with numerous interacting variables. ML models may not always capture the full scope of these interactions, potentially leading to oversimplification or missed nuances in predictions. Therefore, domain knowledge is critical to informing the development of these models. In addition, ML models are optimized based on the data they are trained on, and their predictions can become unreliable if exposed to input data that fall outside of this range. In the context of California's dynamic water and environmental systems, this could occur when

models are applied to scenarios or conditions not well represented in the training data, such as extreme weather events or changes in operations. Using ML models in such situations may lead to inaccurate or misleading predictions, undermining their utility for decision making. Moreover, California's water and environmental policies are highly regulated, with strict ethical and regulatory considerations. ML models must comply with these regulations, particularly regarding data privacy and equity. Finally, ML models, especially those requiring large datasets or complex algorithms, can demand significant computational resources. This can limit some studies and applications, particularly in resource-constrained settings, necessitating careful planning for infrastructure and costs.

5. Conclusions and Future Directions

The protocols presented in this paper provide a step-by-step guide for applying machine learning in water and environmental modeling. While they are informed by the DWR's three decades of relevant experience in California, they are sufficiently general to be applicable to other regions beyond the State. The core of the protocols lies in careful pre-modeling considerations. This includes clearly defining the water and environmental problems one is trying to solve and collecting and preparing high-quality data. The protocols then delve into the model development process, from choosing the most suitable ML architectures for the specific task to train, evaluate, and identify the most desirable models. Finally, the protocols cover the deployment. They outline strategies for integrating the model into real-world applications and effectively communicating its results to ML model users. This paper also includes a case study demonstrating each of these steps.

This paper intends to serve as a dynamic roadmap for harnessing the power of ML in California's water and environmental modeling endeavors. Conceived as a living paper, it will be continuously updated and expanded to address the State's evolving challenges in water resource management and environmental protection.

First, as California grapples with new and ever-shifting water and environmental issues, the protocols will be tailored to tackle emerging threats and opportunities. This might involve incorporating novel data sources, like high-resolution satellite imagery, which are not addressed specifically in the current paper. Recent studies have shown that combining ML with satellite remote sensing data can advance water and environmental modeling. For example, Fleming et al. [36] enhanced seasonal water supply forecasts by integrating satellite-based snow cover data into AI-driven models, while Poon and Kinoshita [60] estimated post-fire evapotranspiration by using ML with satellite-derived land and vegetation indices. Additionally, communication strategies will be honed to effectively reach a wider range of stakeholders, encompassing policymakers, resource managers, and the public.

The second front for advancement lies within the burgeoning field of ML itself. The emergence of groundbreaking techniques like explainable AI and interpretable AI [207], digital twins (DTs) [208,209], and generative artificial intelligence (GenAI) [115] presents immense potential for water and environmental modeling. A critical barrier to the widespread adoption of ML in the operational community is its perceived lack of interpretability. Unlike traditional process-based models, which are grounded in well-established physical principles, many ML models function as black boxes, making it difficult to understand how inputs influence predictions. To address this challenge, researchers have developed various methods to enhance ML interpretability. One of the earliest approaches is graphical sensitivity analysis, which helps visualize how input variables influence model predictions. Cannon and McKendry [210] applied this technique to ANNs for Indian monsoon rainfall prediction. To further bridge the gap between data-driven ML models and physically interpretable scientific models, theory-guided data science (TGDS) has emerged as a promising

approach. Karpatne et al. [211] introduced TGDS as a paradigm that infuses scientific knowledge into ML models, ensuring their predictions align with established physical laws. Building on TGDS, Physics-Informed Neural Networks (PINNs) [112,212–214] and differentiable modeling [215] provide more advanced frameworks for integrating ML with traditional physical models. Shen et al. [187] described differentiable modeling as a method that unifies ML and process-based models through the gradient-based optimization of physically constrained models. PINNs and differentiable geoscientific models are gaining increasing interest and have been successfully applied across various water and environmental studies, improving both interpretability and predictive accuracy (e.g., [83,189]).

DTs present an exciting opportunity to develop more holistic and dynamic representations of environmental systems (e.g., [216,217]). By coupling real-time sensor data with physics-based models and ML algorithms, DTs offer a comprehensive platform for simulating and optimizing environmental processes, thereby facilitating proactive and data-driven decision making in water resource management and environmental conservation efforts. The protocols presented in this paper are expected to be expanded in the future to include guidance on interpretable ML models, as well as the creation and utilization of DTs specifically tailored for water and environmental modeling in California.

GenAI has the potential to revolutionize environmental modeling by enabling the synthesis of realistic and diverse environmental scenarios, thus enhancing the robustness and adaptability of our models to unforeseen circumstances and novel challenges [115,218–220]. Looking ahead, GenAI can play a transformative role in water and environmental management in California, though these applications remain conceptual for now. One possibility is leveraging GenAI-powered chatbots as intelligent knowledge hubs, providing answers to inquiries about the State's water resources, environmental issues, and related topics. These could include general information, historical data, scientific studies, regulations, policies, and legal frameworks—essentially functioning as advanced, domain-specific search engines. Such tools could support research, training, and public outreach efforts. Additionally, GenAI could be used to enhance water and environmental modeling by seamlessly integrating mechanistic and machine learning models to improve predictive accuracy. It might also assist in dynamically selecting the most suitable model for a given task. Another potential application involves deploying GenAI-driven agents for autonomous decision support. These agents could, for instance, process hydro-meteorological forecasts, use them to run hydrologic models for streamflow forecasting, and apply forecasted streamflow to guide reservoir operations or other operational decisions. In 2023, the Governor of California issued Executive Order (EO) N-12-23 to explore the development, application, and associated risks of AI technology, particularly GenAI, across the State. This initiative aims to establish a cautious and responsible approach to assessing and implementing GenAI within state governance, recognizing its significant impact. The protocols outlined in the current paper are not tailored specifically for GenAI. However, the guidelines and best practices on GenAI resulting from EO N-12-23 will be adapted and integrated into future iterations of these protocols.

Finally, ML models, particularly deep learning models for complex water and environmental systems like those in California, demand significant computing resources. While quantum computing has been explored for decades [221,222], recent breakthroughs [223] have unlocked its potential to revolutionize ML applications across various fields, including water and environmental modeling. Unlike classical computing, which processes data in binary digits (0s and 1s), quantum computing uses quantum bits (qubits), which can exist in multiple states simultaneously. This capability allows quantum computers to solve complex problems, process vast datasets, and perform high-dimensional calculations far more efficiently than traditional computers. In water and environmental modeling,

quantum-enhanced ML could improve the accuracy and speed of simulations involving complex hydrological systems, optimize resource allocation, and refine predictive models for climate change impacts, droughts, water quality, and others. Additionally, quantum algorithms could enable the integration of diverse data sources—such as satellite imagery, sensor networks, and historical records—into unified models that better capture California's complex water and environmental systems. While the technology is still in its early stages, ongoing advancements suggest that quantum computing could become a transformative tool for addressing California's water and environmental challenges in the coming decades. This tool will be explored in the near future and its findings will be reported in subsequent updates to the protocols.

Author Contributions: Conceptualization, P.S. and M.H.; methodology, M.H.; software, P.N.; validation, M.H. and P.S.; formal analysis, M.H.; investigation, M.H. and P.N.; resources, E.R.; data curation, P.N.; writing—original draft preparation, M.H.; writing—review and editing, P.S., P.N., E.R., K.G. and F.C.; visualization, M.H. and P.N.; supervision, P.S. and E.R.; project administration, P.N. and E.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All data used in this study are available at https://github.com/CADWRDeltaModeling/Ion_Study_Dashboard/tree/main (accessed on 1 March 2024).

Acknowledgments: The authors would like to sincerely thank their colleagues and collaborators for their invaluable dedication over the past three decades to advancing the development and application of machine learning models in water and environmental modeling in California. This work would not have been possible without their vision, passion, enthusiasm, and unwavering contributions to this field. We would like to express our sincere appreciation to the three anonymous reviewers and the guest editors for their thoughtful and constructive feedback. Their deep knowledge of machine learning and insightful comments were instrumental in improving the overall quality of this work. During the preparation of this work, the authors used ChatGPT to check the grammar and improve the language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the final content of the article. The views expressed in this paper are those of the authors, and not of the State of California.

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

Appendix A. Artificial Intelligence and Machine Learning

Artificial intelligence (AI) is a specific field within the broader discipline of computer science (CS) (Figure A1). It is an umbrella term encompassing the development of intelligent systems capable of mimicking human cognitive functions like learning, problem solving, and decision making [223]. The history of AI dates back to the mid-20th century, with early pioneers like Alan Turing laying the foundation for the field with his seminal work on the Turing test [224], a theoretical test of a machine's ability to exhibit intelligent behavior equivalent to or indistinguishable from that of a human. AI has undergone a remarkable evolution since its inception, showcasing diverse applications across various fields. For example, in engineering, AI optimizes designs, streamlines production processes, and enhances product performance through techniques like predictive maintenance and generative design. As another example, in scientific research, AI facilitates data analysis, accelerates discoveries, and models complex systems across disciplines such as physics, chemistry, and biology.

Machine learning (ML) is a specific subfield of AI that focuses on developing algorithms and techniques that allow computers to learn from data without explicit programming (Figure A1) [183]. This learning process typically involves exposing the algorithm to a large dataset, enabling it to identify patterns and relationships within the data. Over

time, the algorithm improves its performance on specific tasks, such as image recognition, speech recognition, or predicting future outcomes. The history of ML is deeply intertwined with the development of AI, with key advancements in computing power and statistical methods facilitating the creation of powerful ML algorithms.

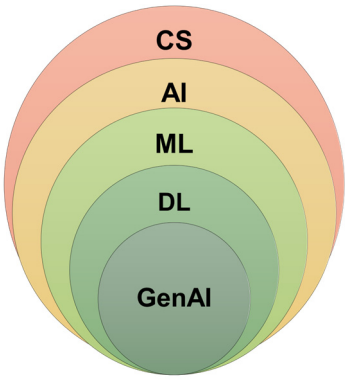

Figure A1. Schematic illustrating relationship among computer science (CS), artificial intelligence (AI), machine learning (ML), deep learning (DL), and generative AI (GenAI).

While AI encompasses a broader range of techniques and goals, including symbolic reasoning and expert systems, ML serves as a crucial engine for enabling machines to learn and improve their performance on specific tasks [225,226]. It is important to understand that not all AI applications rely on ML, but the vast majority of modern AI advancements leverage the power of ML algorithms to achieve their goals. In essence, AI defines the "what"—the broad objective of creating intelligent machines—while ML provides the "how"—the specific techniques and algorithms that enable machines to learn and improve.

Deep learning (DL) is a subset of ML (Figure 1) [227,228]. While ML encompasses a broad spectrum of algorithms and techniques aimed at enabling computers to learn from data and make predictions or decisions, DL specifically focuses on using Artificial Neural Networks (ANNs) [229] with multiple layers to model and understand complex patterns. Unlike traditional ML methods that rely on human experts to define input features, DL algorithms can autonomously learn hierarchical representations of data directly from raw inputs. This ability to automatically extract intricate features from large datasets has led to breakthroughs in various domains, such as image recognition, natural language processing, and speech recognition. DL models, such as Convolutional Neural Networks (CNNs) [230] for images and Recurrent Neural Networks (RNNs) [231] for sequential data, have demonstrated unparalleled performance in tasks ranging from image classification to language translation, revolutionizing industries and driving advancements in AI research.

Generative AI (GenAI) is a subfield of DL (Figure A1) focused on creating entirely new data, rather than just making predictions based on existing information [115]. DL plays a pivotal role in generative AI by providing powerful frameworks, such as generative adversarial networks (GANs) [232] and variational autoencoders (VAEs) [233]. These models are trained on a massive number of data, allowing them to learn the underlying structure and relationships within those data. Once trained, GenAI can use this knowledge to produce entirely new outputs that closely resemble but are not copies of the data it studied. This opens doors for applications like generating text, images, speeches, and videos. Essentially, GenAI utilizes the power of DL to move beyond analysis and toward content creation.

Appendix B. Terms and Definitions

Term	Definition
Accuracy	Correctness of predictions made by a machine learning model.
Activation function	A mathematical operation applied to a neuron's output to introduce non-linearity, enabling the model to learn complex relationships between inputs and outputs.
Adaptability	The capability of a model to adjust and perform effectively in new or changing environments or tasks.
Artificial intelligence (AI)	The development of computer systems that can perform tasks that typically require human intelligence.
Artificial Neural Network (ANN)	A model inspired by the structure and function of biological neural networks, consisting of interconnected nodes (neurons) organized into layers for learning from data.
Bagging	An ensemble learning technique that improves model stability and accuracy by training multiple models on different subsets of the training data and averaging their predictions (for regression) or using majority voting (for classification).
Boosting	An ensemble learning technique that combines multiple weak learners to create a strong learner by iteratively correcting the errors of previous models.
Classification	Techniques used to categorize data into predefined classes or categories based on input features.
Cloud	A network of remote servers hosted on the internet for storing, managing, and processing data and algorithms, providing scalable resources and services.
Communication	The process of conveying updates, status reports, or instructions among team members, stakeholders, and deployed systems to ensure smooth operation and maintenance.
Computational resources	The hardware and software components utilized for training, inference, and executing machine learning algorithms, including CPUs, GPUs, and memory.
Computer science (CS)	The study of the theory, design, and implementation of computer systems and algorithms, including hardware, software, and networking, to solve problems and develop innovative technologies.
Cross-validation	A technique used to assess the performance and generalization ability of a model by dividing the dataset into multiple subsets for training and testing.
Data augmentation	A regularization technique used to artificially increase the size and diversity of a training dataset by applying transformations to existing data.
Data collection	The gathering of relevant information or samples, often from various sources, to build a dataset suitable for training and evaluating machine learning models.
Data pre-processing	The processing of cleaning, transforming, and preparing raw data to make them suitable for analysis and machine learning model training.
Deep learning (DL)	A subset of machine learning where Artificial Neural Networks with multiple layers learn to represent data in increasingly abstract and complex ways.
Domain expertise	Specialized knowledge and understanding of a particular subject area or industry that informs the development and application of machine learning models within that domain.
Dropout	A regularization technique used in neural networks to prevent overfitting by randomly setting a fraction of the neurons (or units) to zero during training.
Early stopping	A technique used in machine learning to prevent overfitting during the training of a model by stopping the training process before the model has fully converged if its performance on a validation set stops improving.

Term	Definition
Ensemble	A technique where multiple models are combined to improve predictive performance by aggregating their individual predictions.
Evaluation	The process of assessing the performance, robustness, and effectiveness of a trained machine learning model by using various metrics and techniques.
Explainability	The degree to which the inner workings and decisions of a model can be understood and interpreted by humans.
Explainable AI	The development of models whose internal logic and reasoning can be transparently understood by human experts.
Features	The data attributes or variables that are fed into a machine learning model for training or making predictions.
Generalization	A model's capacity to effectively learn from training data and apply that knowledge to accurately predict outcomes for unseen data.
Generative AI (GenAI)	A subset of deep learning that generates new content based on patterns learned from existing data.
Gradient	The vector of partial derivatives of a function with respect to its input variables.
Grid search	An exhaustive hyperparameter optimization technique that systematically evaluates a large number of possible combinations of a predefined set of hyperparameters to find the best-performing model configuration.
Hidden layer	An intermediate layer of neurons between the input and output layers in a neural network responsible for extracting and transforming features from the input data.
Hyperband	A hyperparameter optimization algorithm that combines successive halving and bandit-based strategies to allocate computational resources dynamically, prioritizing the most promising configurations while pruning less effective ones.
Hyperparameter	A configuration parameter external to the model that influences its learning process and performance, typically set before training.
Interoperability	The ability of different models or systems to seamlessly exchange and utilize data or functionality, promoting integration and collaboration across diverse environments.
Interpretability	The degree to which a model's internal decision-making process can be understood by humans.
Learning algorithm	A set of procedures and rules that a model follows to adjust its parameters based on input data, aiming to minimize a predefined loss or error function.
Learning rate	A hyperparameter that controls how quickly a model adapts to new information, determining the step size for each iteration's weight updates, balancing convergence speed and accuracy.
Loss function	A function quantifying the difference between predicted and actual values, guiding the optimization process during model training.
Machine learning (ML)	The field of study that enables computers to learn data without being explicitly programmed.
Model architecture	The high-level design of a model, specifying the organization and connection of its components for data processing and transformation.
Model deployment	The process of integrating a trained, tested, and evaluated machine learning model into a production environment.
Model selection	The process of choosing the most appropriate machine learning algorithm or architecture for a given task by comparing and evaluating multiple candidates.

Term	Definition
Neuron	Artificial Neural Network component that processes inputs, applies weights, and produces outputs, mimicking biological neurons' behavior, to learn and make predictions or decisions.
Noise	Irrelevant or random fluctuations in data that can interfere with the learning process or affect the accuracy of a model's predictions.
Normalization	The process of rescaling input features to a predetermined range, to ensure consistent scales and improve convergence in training.
Open source	Software made available under a license that allows users to freely access, modify, and distribute the source code, promoting collaboration, community-driven development, and transparent innovation.
Overfitting	A model's tendency to capture noise or random fluctuations in the training data, resulting in poor performance on unseen data.
Privacy	Safeguarding sensitive information and preserving the confidentiality of data used in models, preventing unauthorized access or disclosure.
Problem definition	Articulating the task or objective that the machine learning model aims to solve.
Protocols	Standardized procedures and guidelines for building, training, and evaluating machine learning models.
Pruning	A regularization technique used to reduce the size of a machine learning model by removing less important parameters (such as weights, neurons, or branches), thereby improving efficiency while maintaining performance.
Quality assurance	Ensuring that products or services meet specified requirements and standards through systematic processes and testing procedures.
Regression	Technique to establish a relationship between independent variables and a dependent variable for prediction.
Regularization	Technique applied during model training to prevent overfitting by penalizing overly complex models.
Reliability	A model's ability to produce consistent and accurate results over time.
Reproducibility	The ability to repeat and obtain consistent results in an experiment, study, or computation.
Robustness	A model's ability to maintain performance when faced with variations in the data or unexpected conditions.
Scalability	The ability of a machine learning model to handle increasing quantities of data, computation, or users while maintaining performance and efficiency.
Security	Protecting models, data, and systems from unauthorized access, manipulation, or adversarial attacks, ensuring confidentiality, integrity, and availability.
Sensitivity	A measure of how much a model's output changes in response to small changes in its inputs or parameters.
Stacking	An ensemble learning technique that combines multiple base models (often of different types) by training a meta-model (or "blender") to make final predictions based on the outputs of those base models.
Supervised learning	The paradigm where models are trained on labeled data, with input–output pairs provided, to learn the mapping between inputs and outputs.
Standardization	Establishing uniform processes, specifications, or data formats to ensure compatibility and efficient operation.
Testing	Evaluating the performance and generalization ability of a trained machine learning model on unseen data to assess its accuracy and reliability.

Term	Definition
Training	The process of teaching a machine learning model to recognize patterns and make predictions by exposing it to data and adjusting its parameters through iterative optimization algorithms.
Transparency	The quality of being clear, open, and understandable.
Uncertainty	The lack of confidence or variability in predictions made by a model.
Unsupervised learning	The paradigm where models are trained on unlabeled data to discover patterns, structures, or relationships without explicit guidance on the desired output.

Appendix C. Summary of Machine Learning Applications in California

Table A1. List of studies on machine learning in water and environmental modeling in California.

Discipline	Target	Reference ID
Surface water hydrology	Streamflow	[29,31,32,37–39,41,42,44]
	Soil moisture	[30,33]
	Seasonal runoff volume	[34–36]
	Reservoir inflow/outflow	[40,43]
Groundwater hydrology	Groundwater storage	[45,52]
	Groundwater age	[46]
	Groundwater level	[47–51]
Hydro-meteorology/hydro-Climatology	Reference evapotranspiration	[53,54]
	Precipitation	[55–59,61,62,64]
	Evapotranspiration	[60]
	Snow water equivalent	[63]
Water quality	Water temperature	[65,85,166]
	Salinity	[66,67,70,71,77–80,82,83,170–174]
	Dissolved oxygen	[66]
	Nitrate	[68,81]
	Trichloropropane	[69]
	Trihalomethane	[167]
	Arsenic	[72]
	Sediment	[73]
	Uranium	[74]
	Ion constituents	[75,76]
	Fecal indicator bacteria	[84]

Table A1. *Cont.*

Discipline	Target	Reference ID
Ecology	Wildfire	[86–89,91,93,94]
	Forage	[90]
	Chinook salmon migration	[92]
	Kelp biodiversity	[95]
	Fish biomass	[95]
	Rocky intertidal biodiversity	[95]
	Fish entrainment	[96]
	Almond yield	[97]
	Crop	[98]
	Harmful algal bloom	[178]
Water management/operations	Community water systems	[19]
	Agricultural water shortage	[99]
	Flood	[100,104,107]
	Irrigation water demand	[101]
	Water security/access	[102]
	Drought	[103]
	Reservoir operations	[105,108]
	Water supply vulnerability	[106]
	Water use pattern	[109]
	Marginal export cost	[168,169,175–177]

Appendix D. Supplementary Section for the Case Study

Table A2. Overview of key features of selected machine learning models in the case study *.

Feature/Model	Decision Trees	Random Forest	Gradient Boosting	Artificial Neural Networks
Model type	Tree-based	Ensemble	Ensemble	Neural network
Basic unit	Decision Trees	Decision Trees	Weak learners	Neurons
Hidden layers	N/A	N/A	N/A	One or more
Loss function	Gini/Entropy	Gini/Entropy	Various	MSE, Cross-Entropy, etc.
Learning algorithm	ID3, CART, etc.	Bagging	Boosting	Gradient Descent, Adam, etc.
Regularization	Pruning	Voting/averaging	Shrinkage	Dropout, Weight Decay, etc.
Scalability	Moderate	High	Moderate to high	High
Robustness	Moderate	High	High	Varies
Interpretability	High	Moderate	Low	Low
Speed/efficiency (training)	Fast	Moderate	Moderate	Varies

Table A2. *Cont.*

Feature/Model	Decision Trees	Random Forest	Gradient Boosting	Artificial Neural Networks
Speed/efficiency (inference)	Fast	Fast	Fast	Fast
Applications	Classification and regression	Classification, regression, and anomaly detection	Classification, regression, and ranking	Classification, regression, NLP, and image processing

* Adapted from [76].

Table A3. Optimized hyperparameters for each ion constituent ANN model *.

Hidden Layer	TDSs		Mg^{2+}		Na^{+}	
	N	Act	N	Act	N	Act
1	30	elu	30	relu	30	tanh
2	30	sigmoid	30	elu	30	elu
3	30	elu	30	tanh	30	sigmoid
4	30	relu	30	relu	30	elu
Hidden Layer	**Ca^{2+}**		**Cl^{-}**		**SO_4^{2-}**	
	N	**Act**	**N**	**Act**	**N**	**Act**
1	40	elu	30	relu	44	relu
2	40	sigmoid	30	elu	44	relu
3	40	relu	30	sigmoid	44	relu
4	30	tanh	30	elu	22	relu
Hidden Layer	**Br^{-}**		**Alkalinity**		**K^{+}**	
	N	**Act**	**N**	**Act**	**N**	**Act**
1	44	elu	30	tanh	44	relu
2	44	sigmoid	30	relu	44	relu
3	30	elu	30	tanh	44	relu
4	30	tanh	30	elu	22	relu

* Adapted from [76].

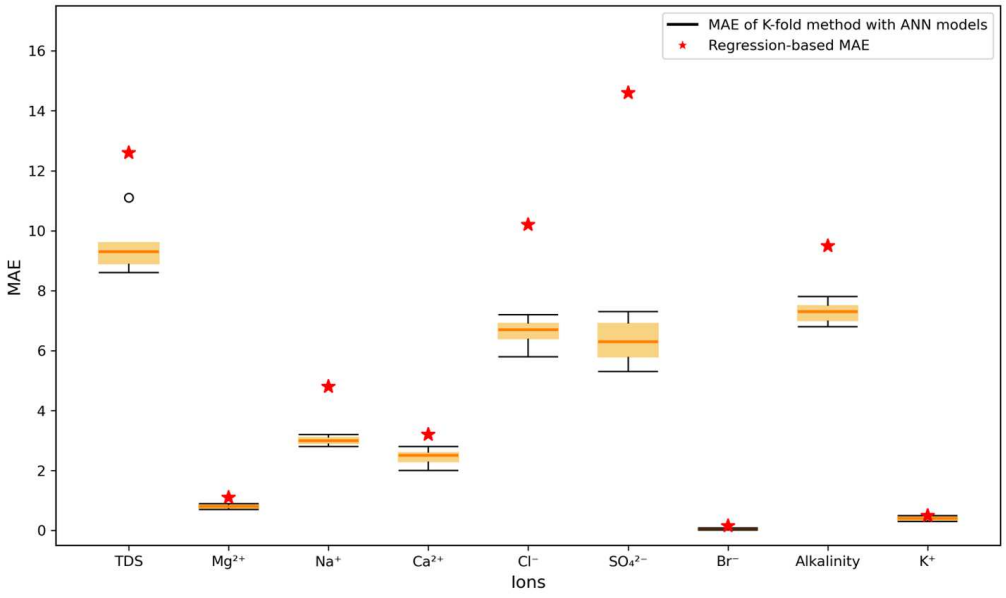

Figure A2. Box plots of the Mean Absolute Error (MAE) values of nine ion constituents across the 5-fold cross-validation using ANN models versus the corresponding MAE values of the benchmark regression model (adapted from [76]).

References

1. Pinter, N.; Lund, J.; Moyle, P. The California Water Model: Resilience through failure. *Hydrol. Process.* **2019**, *33*, 1775–1779. [CrossRef]
2. Hanak, E.; Lund, J.R. Adapting California's water management to climate change. *Clim. Chang.* **2012**, *111*, 17–44. [CrossRef]
3. Bay-Delta Modeling Forum (BDMF). *Protocols for Water and Environmental Modeling*; California Water and Environmental Modeling Forum: Sacramento, CA, USA, 2000; p. 47.
4. California Water & Environmental Modeling Forum (CWEMF). *Protocols for Water and Environmental Modeling*; California Water and Environmental Modeling Forum: Sacramento, CA, USA, 2021; p. 119.
5. Hannaford, J.F. Multiple-graphical correlation for water supply forecasting. In Proceedings of the 24th Annual Western Snow Conference, Penticton, BC, Canada, 17–19 April 1956.
6. Hart, D.; Gehrke, F. Status of the California cooperative snow survey program. In Proceedings of the 58th Annual Western Snow Conference, Sacramento, CA, USA, 17–19 April 1990.
7. California Department of Water Resources (DWR; formerly the Division of Water Resources). *Variation and Control of Salinity*; California Division of Water Resources Bulletin: Sacramento, CA, USA, 1931.
8. California Department of Water Resources. Dayflow Documentation [Internet]. Available online: https://water.ca.gov/Programs/Integrated-Science-and-Engineering/Compliance-Monitoring-And-Assessment/Dayflow-Data (accessed on 1 May 2024).
9. Becker, L.; Sparks, D.; Fults, D.M.; Yeh, W.W. Operations models for central valley project. *J. Water Resour. Plan. Manag. Div.* **1976**, *102*, 101–115. [CrossRef]
10. Draper, A.J.; Munévar, A.; Arora, S.K.; Reyes, E.; Parker, N.L.; Chung, F.I.; Peterson, L.E. CalSim: Generalized model for reservoir system analysis. *J. Water Resour. Plan. Manag.* **2004**, *130*, 480–489. [CrossRef]
11. California Department of Water Resources (DWR). Calibration and verification of DWRDSM. In *Methodology for Flow and Salinity Estimates in the Sacramento-San Joaquin Delta and Suisun Marsh: 12th Annual Progress Report*; California Department of Water Resources: Sacramento, CA, USA, 1991.
12. Demargne, J.; Wu, L.; Regonda, S.K.; Brown, J.D.; Lee, H.; He, M.; Seo, D.J.; Hartman, R.; Herr, H.D.; Fresch, M.; et al. The science of NOAA's operational hydrologic ensemble forecast service. *Bull. Am. Meteorol. Soc.* **2014**, *95*, 79–98. [CrossRef]
13. Brown, J.D.; Wu, L.; He, M.; Regonda, S.; Lee, H.; Seo, D.J. Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 1. Experimental design and forcing verification. *J. Hydrol.* **2014**, *519*, 2869–2889. [CrossRef]
14. He, M.; Whitin, B.; Hartman, R.; Henkel, A.; Fickenschers, P.; Staggs, S.; Morin, A.; Imgarten, M.; Haynes, A.; Russo, M. Verification of Ensemble Water Supply Forecasts for Sierra Nevada Watersheds. *Hydrology* **2016**, *3*, 35. [CrossRef]
15. Brown, J.D.; He, M.; Regonda, S.; Wu, L.; Lee, H.; Seo, D.J. Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 2. Streamflow verification. *J. Hydrol.* **2014**, *519*, 2847–2868. [CrossRef]
16. Rosenberg, E.A.; Wood, A.W.; Steinemann, A.C. Statistical applications of physically based hydrologic models to seasonal stream-flow forecasts. *Water Resour. Res.* **2011**, *47*. [CrossRef]
17. He, M.; Russo, M.; Anderson, M. Predictability of seasonal streamflow in a changing climate in the Sierra Nevada. *Climate* **2016**, *4*, 57. [CrossRef]
18. Shukla, S.; Safeeq, M.; AghaKouchak, A.; Guan, K.; Funk, C. Temperature impacts on the water year 2014 drought in California. *Geophys. Res. Lett.* **2015**, *42*, 4384–4393. [CrossRef]
19. Wang, D.; Bui, H.; Zúñiga, E.J. How Machine Learning Can Help Water Systems During Drought. *J. Am. Water Work. Assoc.* **2024**, *116*, 32–43. [CrossRef]
20. Allaire, M.; Acquah, S. Disparities in drinking water compliance: Implications for incorporating equity into regulatory practices. *AWWA Water Sci.* **2022**, *4*, e1274. [CrossRef]
21. Schwarz, A.M. California Central Valley water rights in a changing climate. *San Franc. Estuary Watershed Sci.* **2015**, *13*, 1–11. [CrossRef]
22. Arnold, W.; Spanos, K.A.; Schwarz, A.; Doan, C.; Deaver, K.; Akens, M.U.; Maendly, R.; Selmon, M.; Ekstrom, J.; Coombe, P.; et al. A Climate Action Plan for the California Department of Water Resources. *J. Am. Water Work. Assoc.* **2022**, *114*, 10–20. [CrossRef]
23. Ray, P.; Wi, S.; Schwarz, A.; Correa, M.; He, M.; Brown, C. Vulnerability and risk: Climate change and water supply from California's Central Valley water system. *Clim. Chang.* **2020**, *161*, 177–199. [CrossRef]
24. Hamilton, A.L.; Zeff, H.B.; Characklis, G.W.; Reed, P.M. Resilient California water portfolios require infrastructure investment partnerships that are viable for all partners. *Earth's Futur.* **2022**, *10*, e2021EF002573. [CrossRef]
25. DiGennaro, B.; Reed, D.; Swanson, C.; Hastings, L.; Hymanson, Z.; Healey, M.; Siegel, S.; Cantrell, S.; Herbold, B. Using conceptual models in ecosystem restoration decision making: An example from the Sacramento-San Joaquin River Delta, California. *San Franc. Estuary Watershed Sci.* **2012**, *10*, 1–15. [CrossRef]

26. MacWilliams, M.L.; Ateljevich, E.S.; Monismith, S.G.; Enright, C. An overview of multi-dimensional models of the Sacramento–San Joaquin Delta. *San Franc. Estuary Watershed Sci.* **2016**, *14*, 1–35. [CrossRef]
27. Sommer, T.; Hartman, R.; Koller, M.; Koohafkan, M.; Conrad, J.L.; MacWilliams, M.; Bever, A.; Burdi, C.; Hennessy, A.; Beakes, M. Evaluation of a large-scale flow manipulation to the upper San Francisco Estuary: Response of habitat conditions for an en-dangered native fish. *PLoS ONE* **2020**, *15*, e0234673. [CrossRef]
28. Draper, A.J.; Jenkins, M.W.; Kirby, K.W.; Lund, J.R.; Howitt, R.E. Economic-engineering optimization for California water management. *J. Water Resour. Plan. Manag.* **2003**, *129*, 155–164. [CrossRef]
29. Adera, S.; Bellugi, D.; Dhakal, A.; Larsen, L. Streamflow prediction at the intersection of physics and machine learning: A case study of two Mediterranean-climate watersheds. *Water Resour. Res.* **2024**, *60*, e2023WR035790. [CrossRef]
30. Basak, A.; Schmidt, K.M.; Mengshoel, O.J. From data to interpretable models: Machine learning for soil moisture forecasting. *Int. J. Data Sci. Anal.* **2023**, *15*, 9–32. [CrossRef] [PubMed]
31. Duan, S.; Pallotta, G.; Bonfils, C. Higher-order internal modes of variability imprinted in year-to-year California streamflow changes. *Commun. Earth Environ.* **2024**, *5*, 448. [CrossRef]
32. Duan, S.; Ullrich, P.; Shu, L. Using convolutional neural networks for streamflow projection in California. *Front. Water* **2020**, *2*, 28. [CrossRef]
33. Fang, K.; Shen, C. Near-real-time forecast of satellite-based soil moisture using long short-term memory with an adaptive data integration kernel. *J. Hydrometeorol.* **2020**, *21*, 399–413. [CrossRef]
34. Fleming, S.W.; Goodbody, A.G. A machine learning metasystem for robust probabilistic nonlinear regression-based forecasting of seasonal water availability in the US west. *IEEE Access* **2019**, *7*, 119943–119964. [CrossRef]
35. Fleming, S.W.; Garen, D.C.; Goodbody, A.G.; McCarthy, C.S.; Landers, L.C. Assessing the new Natural Resources Conservation Service water supply forecast model for the American West: A challenging test of explainable, automated, ensemble artificial intelligence. *J. Hydrol.* **2021**, *602*, 126782. [CrossRef]
36. Fleming, S.W.; Rittger, K.; Oaida Taglialatela, C.M.; Graczyk, I. Leveraging next-generation satellite remote sensing-based snow data to improve seasonal water supply predictions in a practical machine learning-driven river forecast system. *Water Resour. Res.* **2024**, *60*, e2023WR035785. [CrossRef]
37. Grantham, T.E.; Zimmerman, J.K.; Carah, J.K.; Howard, J.K. Stream flow modeling tools inform environmental water policy in California. *Calif. Agric.* **2019**, *73*, 33–39. [CrossRef]
38. Kratzert, F.; Klotz, D.; Herrnegger, M.; Sampson, A.K.; Hochreiter, S.; Nearing, G.S. Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resour. Res.* **2019**, *55*, 11344–11354. [CrossRef]
39. Kratzert, F.; Klotz, D.; Shalev, G.; Klambauer, G.; Hochreiter, S.; Nearing, G. Benchmarking a catchment-aware long short-term memory network (LSTM) for large-scale hydrological modeling. *Hydrol. Earth Syst. Sci. Discuss* **2019**, *2019*, 1–32.
40. Maddu, R.; Pradhan, I.; Ahmadisharaf, E.; Singh, S.K.; Shaik, R. Short-range reservoir inflow forecasting using hydrological and large-scale atmospheric circulation information. *J. Hydrol.* **2022**, *612*, 128153. [CrossRef]
41. Thakur, B.; Kalra, A.; Ahmad, S.; Lamb, K.W.; Lakshmi, V. Bringing statistical learning machines together for hydro-climatological predictions-Case study for Sacramento San-Joaquin River Basin, California. *J. Hydrol. Reg. Stud.* **2020**, *27*, 100651. [CrossRef]
42. Wi, S.; Steinschneider, S. Assessing the physical realism of deep learning hydrologic model projections under climate change. *Water Resour. Res.* **2022**, *58*, e2022WR032123. [CrossRef]
43. Yang, T.; Asanjan, A.A.; Welles, E.; Gao, X.; Sorooshian, S.; Liu, X. Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. *Water Resour. Res.* **2017**, *53*, 2786–2812. [CrossRef]
44. Zimmerman, J.K.; Carlisle, D.M.; May, J.T.; Klausmeyer, K.R.; Grantham, T.E.; Brown, L.R.; Howard, J.K. Patterns and magnitude of flow alteration in California, USA. *Freshw. Biol.* **2018**, *63*, 859–873. [CrossRef]
45. Agarwal, V.; Akyilmaz, O.; Shum, C.K.; Feng, W.; Yang, T.-Y.; Forootan, E.; Syed, T.H.; Haritashya, U.K.; Uz, M. Machine learning based downscaling of GRACE-estimated groundwater in Central Valley, California. *Sci. Total Environ.* **2023**, *865*, 161138. [CrossRef]
46. Azhar, A.; Chakraborty, I.; Visser, A.; Liu, Y.; Lerback, J.C.; Oerter, E. Machine learning prediction of tritium-helium groundwater ages in the Central Valley, California, USA. *Water Resour. Res.* **2025**, *61*, e2024WR038031. [CrossRef]
47. May-Lagunes, G.; Chau, V.; Ellestad, E.; Greengard, L.; D'Odorico, P.; Vahabi, P.; Todeschini, A.; Girotto, M. Forecasting groundwater levels using machine learning methods: The case of California's Central Valley. *J. Hydrol. X* **2023**, *21*, 100161. [CrossRef]
48. Miro, M.E.; Groves, D.; Tincher, B.; Syme, J.; Tanverakul, S.; Catt, D. Adaptive water management in the face of uncertainty: Integrating machine learning, groundwater modeling and robust decision making. *Clim. Risk Manag.* **2021**, *34*, 100383. [CrossRef]
49. Müller, J.; Park, J.; Sahu, R.; Varadharajan, C.; Arora, B.; Faybishenko, B.; Agarwal, D. Surrogate optimization of deep neural networks for groundwater predictions. *J. Glob. Optim.* **2021**, *81*, 203–231. [CrossRef]

50. Park, J.; Müller, J.; Arora, B.; Faybishenko, B.; Pastorello, G.; Varadharajan, C.; Sahu, R.; Agarwal, D. Long-term missing value imputation for time series data using deep neural networks. *Neural Comput. Appl.* **2023**, *35*, 9071–9091. [CrossRef]
51. Rohde, M.M.; Biswas, T.; Housman, I.W.; Campbell, L.S.; Klausmeyer, K.R.; Howard, J.K. A machine learning approach to predict groundwater levels in California reveals ecosystems at risk. *Front. Earth Sci.* **2021**, *9*, 784499. [CrossRef]
52. Yin, J.; Medellín-Azuara, J.; Escriva-Bou, A.; Liu, Z. Bayesian machine learning ensemble approach to quantify model uncertainty in predicting groundwater storage change. *Sci. Total Environ.* **2021**, *769*, 144715. [CrossRef]
53. Ahmadi, A.; Daccache, A.; Sadegh, M.; Snyder, R.L. Statistical and deep learning models for reference evapotranspiration time series forecasting: A comparison of accuracy, complexity, and data efficiency. *Comput. Electron. Agric.* **2023**, *215*, 108424. [CrossRef]
54. Ahmadi, A.; Kazemi, M.H.; Daccache, A.; Snyder, R.L. SolarET: A generalizable machine learning approach to estimate reference evapotranspiration from solar radiation. *Agric. Water Manag.* **2024**, *295*, 108779. [CrossRef]
55. Ghazvinian, M.; Delle Monache, L.; Afzali Gorooh, V.; Steinhoff, D.; Sengupta, A.; Hu, W.; Simpson, M.; Weihs, R.; Papadopoulos, C.; Mulrooney, P.; et al. Deep Learning of a 200-Member Ensemble with a Limited Historical Training to Improve the Prediction of Extreme Precipitation Events. *Mon. Weather. Rev.* **2024**, *152*, 1587–1605. [CrossRef]
56. Ghazvinian, M.; Zhang, Y.; Seo, D.-J.; He, M.; Fernando, N. A novel hybrid artificial neural network-parametric scheme for postprocessing medium-range precipitation forecasts. *Adv. Water Resour.* **2021**, *151*, 103907. [CrossRef]
57. Gibson, P.B.; Chapman, W.E.; Altinok, A.; Delle Monache, L.; DeFlorio, M.J.; Waliser, D.E. Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts. *Commun. Earth Environ.* **2021**, *2*, 159. [CrossRef]
58. Hu, W.; Ghazvinian, M.; Chapman, W.E.; Sengupta, A.; Ralph, F.M.; Delle Monache, L. Deep learning forecast uncertainty for precipitation over the Western United States. *Mon. Weather. Rev.* **2023**, *151*, 1367–1385. [CrossRef]
59. Majnooni, S.; Nikoo, M.R.; Nematollahi, B.; Fooladi, M.; Alamdari, N.; Al-Rawas, G.; Gandomi, A.H. Long-term precipitation prediction in different climate divisions of California using remotely sensed data and machine learning. *Hydrol. Sci. J.* **2023**, *68*, 1984–2008. [CrossRef]
60. Poon, P.K.; Kinoshita, A.M. Estimating evapotranspiration in a post-fire environment using remote sensing and machine learning. *Remote Sens.* **2018**, *10*, 1728. [CrossRef]
61. Scheuerer, M.; Switanek, M.B.; Worsnop, R.P.; Hamill, T.M. Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over California. *Mon. Weather. Rev.* **2020**, *148*, 3489–3506. [CrossRef]
62. Silverman, D.; Dracup, J.A. Artificial neural networks and long-range precipitation prediction in California. *J. Appl. Meteorol.* **2000**, *39*, 57–66. [CrossRef]
63. Song, Y.; Tsai, W.-P.; Gluck, J.; Rhoades, A.; Zarzycki, C.; McCrary, R.; Lawson, K.; Shen, C. LSTM-based data integration to improve snow water equivalent prediction and diagnose error sources. *J. Hydrometeorol.* **2024**, *25*, 223–237. [CrossRef]
64. Zhang, C.; Brodeur, Z.P.; Steinschneider, S.; Herman, J.D. Leveraging spatial patterns in precipitation forecasts using deep learning to support regional water management. *Water Resour. Res.* **2022**, *58*, e2021WR031910. [CrossRef]
65. Abdi, R.; Rust, A.; Hogue, T.S. Development of a multilayer deep neural network model for predicting hourly river water temperature from meteorological data. *Front. Environ. Sci.* **2021**, *9*, 738322. [CrossRef]
66. Chawla, P.; Cao, X.; Fu, Y.; Hu, C.-M.; Wang, M.; Wang, S.; Gao, J.Z. Water quality prediction of Salton sea using machine learning and big data techniques. *Int. J. Environ. Anal. Chem.* **2023**, *103*, 6835–6858. [CrossRef]
67. Chen, L.; Roy, S.B.; Hutton, P.H. Emulation of a process-based estuarine hydrodynamic model. *Hydrol. Sci. J.* **2018**, *63*, 783–802. [CrossRef]
68. Gorski, G.; Fisher, A.T.; Dailey, H.; Beganskas, S.; Schmidt, C. The potential for nitrate removal during infiltration: Mapping with machine learning informed by field and laboratory experiments. *Hydrol. Process.* **2022**, *36*, e14750. [CrossRef]
69. Hauptman, B.H.; Naughton, C.C.; Harmon, T.C. Using machine learning to predict 1, 2, 3-trichloropropane contamination from legacy non-point source pollution of groundwater in California's Central Valley. *Groundw. Sustain. Dev.* **2023**, *22*, 100955. [CrossRef]
70. He, M.; Zhong, L.; Sandhu, P.; Zhou, Y. Emulation of a process-based salinity generator for the Sacramento–San Joaquin Delta of California via deep learning. *Water* **2020**, *12*, 2088. [CrossRef]
71. Jayasundara, N.C.; Seneviratne, S.A.; Reyes, E.; Chung, F.I. Artificial neural network for Sacramento–San Joaquin Delta flow–salinity relationship for CalSim 3.0. *J. Water Resour. Plan. Manag.* **2020**, *146*, 04020015. [CrossRef]
72. Karasaki, S.; Morello-Frosch, R.; Callaway, D. Machine learning for environmental justice: Dissecting an algorithmic approach to predict drinking water quality in California. *Sci. Total Environ.* **2024**, *951*, 175730. [CrossRef] [PubMed]
73. Kim, H.S.; He, M.; Sandhu, P. Suspended sediment concentration estimation in the Sacramento-San Joaquin Delta of California using long short-term memory networks. *Hydrol. Process.* **2022**, *36*, e14694. [CrossRef]
74. Lopez, A.M.; Wells, A.; Fendorf, S. Soil and aquifer properties combine as predictors of groundwater uranium concentrations within the Central Valley, California. *Environ. Sci. Technol.* **2020**, *55*, 352–361. [CrossRef]

75. Namadi, P.; He, M.; Sandhu, P. Salinity-constituent conversion in South Sacramento-San Joaquin Delta of California via machine learning. *Earth Sci. Inform.* **2022**, *15*, 1749–1764. [CrossRef]
76. Namadi, P.; He, M.; Sandhu, P. Modeling ion constituents in the Sacramento-San Joaquin Delta using multiple machine learning approaches. *J. Hydroinform.* **2023**, *25*, 2541–2560. [CrossRef]
77. Qi, S.; Bai, Z.; Ding, Z.; Jayasundara, N.; He, M.; Sandhu, P.; Seneviratne, S.; Kadir, T. Enhanced artificial neural networks for salinity estimation and forecasting in the Sacramento-San Joaquin delta of California. *J. Water Resour. Plan. Manag.* **2021**, *147*, 04021069. [CrossRef]
78. Qi, S.; He, M.; Bai, Z.; Ding, Z.; Sandhu, P.; Chung, F.; Namadi, P.; Zhou, Y.; Hoang, R.; Tom, B.; et al. Novel Salinity Modeling Using Deep Learning for the Sacramento–San Joaquin Delta of California. *Water* **2022**, *14*, 3628. [CrossRef]
79. Qi, S.; He, M.; Bai, Z.; Ding, Z.; Sandhu, P.; Zhou, Y.; Namadi, P.; Tom, B.; Hoang, R.; Anderson, J. Multi-location emulation of a process-based salinity model using machine learning. *Water* **2022**, *14*, 2030. [CrossRef]
80. Qi, S.; He, M.; Hoang, R.; Zhou, Y.; Namadi, P.; Tom, B.; Sandhu, P.; Bai, Z.; Chung, F.; Ding, Z.; et al. Salinity Modeling Using Deep Learning with Data Augmentation and Transfer Learning. *Water* **2023**, *15*, 2482. [CrossRef]
81. Ransom, K.M.; Nolan, B.T.; Traum, J.A.; Faunt, C.C.; Bell, A.M.; Gronberg, J.A.M.; Wheeler, D.C.; Rosecrans, C.Z.; Jurgens, B.; Schwarz, G.E.; et al. A hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley aquifer, California, USA. *Sci. Total Environ.* **2017**, *601*, 1160–1172. [CrossRef] [PubMed]
82. Rath, J.S.; Hutton, P.H.; Chen, L.; Roy, S.B. A hybrid empirical-Bayesian artificial neural network model of salinity in the San Francisco Bay-Delta estuary. *Environ. Model. Softw.* **2017**, *93*, 193–208. [CrossRef]
83. Roh, D.M.; He, M.; Bai, Z.; Sandhu, P.; Chung, F.; Ding, Z.; Qi, S.; Zhou, Y.; Hoang, R.; Namadi, P.; et al. Physics-Informed Neural Networks-Based Salinity Modeling in the Sacramento–San Joaquin Delta of California. *Water* **2023**, *15*, 2320. [CrossRef]
84. Searcy, R.T.; Boehm, A.B. Know before you go: Data-driven beach water quality forecasting. *Environ. Sci. Technol.* **2022**, *57*, 17930–17939. [CrossRef]
85. Wade, J.; Kelleher, C.; Hannah, D.M. Machine learning unravels controls on river water temperature regime dynamics. *J. Hydrol.* **2023**, *623*, 129821. [CrossRef]
86. Bhowmik, R.T.; Jung, Y.S.; Aguilera, J.A.; Prunicki, M.; Nadeau, K. A multi-modal wildfire prediction and early-warning system based on a novel machine learning framework. *J. Environ. Manag.* **2023**, *341*, 117908. [CrossRef]
87. Brown, P.T.; Hanley, H.; Mahesh, A.; Reed, C.; Strenfel, S.J.; Davis, S.J.; Kochanski, A.K.; Clements, C.B. Climate warming increases extreme daily wildfire growth risk in California. *Nature* **2023**, *621*, 760–766. [CrossRef]
88. Chen, A. Evaluating the relationships between wildfires and drought using machine learning. *Int. J. Wildland Fire* **2022**, *31*, 230–239. [CrossRef]
89. Cheng, S.; Jin, Y.; Harrison, S.P.; Quilodrán-Casas, C.; Prentice, I.C.; Guo, Y.K.; Arcucci, R. Parameter flexible wildfire prediction using machine learning techniques: Forward and inverse modelling. *Remote Sens.* **2022**, *14*, 3228. [CrossRef]
90. Liu, H.; Jin, Y.; Roche, L.M.; T O'Geen, A.; Dahlgren, R.A. Understanding spatial variability of forage production in California grasslands: Delineating climate, topography and soil controls. *Environ. Res. Lett.* **2021**, *16*, 014043. [CrossRef]
91. Malik, A.; Rao, M.R.; Puppala, N.; Koouri, P.; Thota, V.A.K.; Liu, Q.; Chiao, S.; Gao, J. Data-driven wildfire risk prediction in northern California. *Atmosphere* **2021**, *12*, 109. [CrossRef]
92. Olivetti, S.; Gil, M.A.; Sridharan, V.K.; Hein, A.M. Merging computational fluid dynamics and machine learning to reveal animal migration strategies. *Methods Ecol. Evol.* **2021**, *12*, 1186–1200. [CrossRef]
93. Qiu, L.; Chen, J.; Fan, L.; Sun, L.; Zheng, C. High-resolution mapping of wildfire drivers in California based on machine learning. *Sci. Total Environ.* **2022**, *833*, 155155. [CrossRef] [PubMed]
94. Reid, C.E.; Jerrett, M.; Petersen, M.L.; Pfister, G.G.; Morefield, P.E.; Tager, I.B.; Raffuse, S.M.; Balmes, J.R. Spatiotemporal prediction of fine particulate matter during the 2008 northern California wildfires using machine learning. *Environ. Sci. Technol.* **2015**, *49*, 3887–3896. [CrossRef]
95. Stock, A.; Haupt, A.J.; Mach, M.E.; Micheli, F. Mapping ecological indicators of human impact with statistical and machine learning methods: Tests on the California coast. *Ecol. Inform.* **2018**, *48*, 37–47. [CrossRef]
96. Tillotson, M.D.; Hassrick, J.; Collins, A.L.; Phillis, C. Machine learning forecasts to reduce risk of entrainment loss of endangered salmonids at large-scale water diversions in the Sacramento–San Joaquin Delta, California. *San Franc. Estuary Watershed Sci.* **2022**, *20*, 1–21. [CrossRef]
97. Zhang, Z.; Jin, Y.; Chen, B.; Brown, P. California almond yield prediction at the orchard level with a machine learning approach. *Front. Plant Sci.* **2019**, *10*, 809. [CrossRef]
98. Zhong, L.; Hu, L.; Zhou, H. Deep learning based multi-temporal crop classification. *Remote Sens. Environ.* **2019**, *221*, 430–443. [CrossRef]
99. Boisramé, G.F.; Carroll, R.W. Exploring Climate-Driven agricultural water shortages in a Snow-Fed basin using a water allocation model and Machine learning. *J. Hydrol.* **2023**, *621*, 129605. [CrossRef]

100. Bowers, C.; Serafin, K.A.; Baker, J.W. Uncovering drivers of atmospheric river flood damage using interpretable machine learning. *Nat. Hazards Rev.* **2024**, *25*, 04024022. [CrossRef]
101. Emami, M.; Ahmadi, A.; Daccache, A.; Nazif, S.; Mousavi, S.-F.; Karami, H. County-level irrigation water demand estimation using machine learning: Case study of California. *Water* **2022**, *14*, 1937. [CrossRef]
102. Gu, Z.; Li, W.; Hanemann, M.; Tsai, Y.; Wutich, A.; Westerhoff, P.; Landes, L.; Roque, A.D.; Zheng, M.; Velasco, C.A.; et al. Applying machine learning to understand water security and water access inequality in underserved colonia communities. *Comput. Environ. Urban Syst.* **2023**, *102*, 101969. [CrossRef]
103. Parajuli, A.; Parajuli, R.; Banjara, M.; Bhusal, A.; Dahal, D.; Kalra, A. Application of Machine Learning and Hydrological Models for Drought Evaluation in Ungauged Basins Using Satellite-Derived Precipitation Data. *Climate* **2024**, *12*, 190. [CrossRef]
104. Rasheed, Z.; Aravamudan, A.; Sefidmazgi, A.G.; Anagnostopoulos, G.C.; Nikolopoulos, E.I. Advancing flood warning procedures in ungauged basins with machine learning. *J. Hydrol.* **2022**, *609*, 127736. [CrossRef]
105. Rieker, J.D.; Labadie, J.W. An intelligent agent for optimal river-reservoir system management. *Water Resour. Res.* **2012**, *48*. [CrossRef]
106. Robinson, B.; Cohen, J.S.; Herman, J.D. Detecting early warning signals of long-term water supply vulnerability using machine learning. *Environ. Model. Softw.* **2020**, *131*, 104781. [CrossRef]
107. Tanim, A.H.; McRae, C.B.; Tavakol-Davani, H.; Goharian, E. Flood detection in urban areas using satellite imagery and machine learning. *Water* **2022**, *14*, 1140. [CrossRef]
108. Yang, T.; Gao, X.; Sorooshian, S.; Li, X. Simulating California reservoir operation using the classification and regression-tree algorithm combined with a shuffled cross-validation scheme. *Water Resour. Res.* **2016**, *52*, 1626–1651. [CrossRef]
109. Zhou, Y.; Khan, B.M.; Choi, J.Y.; Cohen, Y. Machine Learning Modeling of Water Use Patterns in Small Disadvantaged Communities. *Water* **2021**, *13*, 2312. [CrossRef]
110. Breiman, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Stat. Sci.* **2001**, *16*, 199–215. [CrossRef]
111. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; MIT Press: Cambridge, MA, USA, 2012.
112. Karniadakis, G.E.; Kevrekidis, I.G.; Lu, L.; Perdikaris, P.; Wang, S.; Yang, L. Physics-informed machine learning. *Nat. Rev. Phys.* **2021**, *3*, 422–440. [CrossRef]
113. Black, D.C.; Wallbrink, P.J.; Jordan, P.W. Towards best practice implementation and application of models for analysis of water resources management scenarios. *Environ. Model. Softw.* **2014**, *52*, 136–148. [CrossRef]
114. Wu, W.; Dandy, G.C.; Maier, H.R. Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling. *Environ. Model. Softw.* **2014**, *54*, 108–127. [CrossRef]
115. Epstein, Z.; Hertzmann, A. Art and the science of generative AI. *Science* **2023**, *380*, 1110–1111. [CrossRef]
116. McCarthy, J.; Minsky, M.; Rochester, N.; Shannon, C.E. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. August 1955. Available online: http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf (accessed on 1 March 2024).
117. McCulloch, W.S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **1943**, *5*, 115–133. [CrossRef]
118. Zheng, A.; Casari, A. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2018.
119. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
120. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1996**, *58*, 267–288. [CrossRef]
121. Huang, R.; Ma, C.; Ma, J.; Huangfu, X.; He, Q. Machine learning in natural and engineered water systems. *Water Res.* **2021**, *205*, 117666. [CrossRef] [PubMed]
122. Sit, M.; Demiray, B.Z.; Xiang, Z.; Ewing, G.J.; Sermet, Y.; Demir, I. A comprehensive review of deep learning applications in hydrology and water resources. *Water Sci. Technol.* **2020**, *82*, 2635–2670. [CrossRef]
123. Tripathy, K.P.; Mishra, A.K. Deep learning in hydrology and water resources disciplines: Concepts, methods, applications, and research directions. *J. Hydrol.* **2023**, *628*, 130458. [CrossRef]
124. Wai, K.P.; Chia, M.Y.; Koo, C.H.; Huang, Y.F.; Chong, W.C. Applications of deep learning in water quality management: A state-of-the-art review. *J. Hydrol.* **2022**, *613*, 128332. [CrossRef]
125. Chapman, W.E.; Subramanian, A.C.; Delle Monache, L.; Xie, S.P.; Ralph, F.M. Improving atmospheric river forecasts with machine learning. *Geophys. Res. Lett.* **2019**, *46*, 10627–10635. [CrossRef]
126. Chapman, W.E.; Delle Monache, L.; Alessandrini, S.; Subramanian, A.C.; Ralph, F.M.; Xie, S.-P.; Lerch, S.; Hayatbini, N. Probabilistic predictions from deterministic atmospheric river forecasts with deep learning. *Mon. Weather. Rev.* **2022**, *150*, 215–234. [CrossRef]

127. Zarei, M.; Bozorg-Haddad, O.; Baghban, S.; Delpasand, M.; Goharian, E.; Loáiciga, H.A. Machine-learning algorithms for forecast-informed reservoir operation (FIRO) to reduce flood damages. *Sci. Rep.* **2021**, *11*, 24295. [CrossRef]
128. Rajaee, T.; Ebrahimi, H.; Nourani, V. A review of the artificial intelligence methods in groundwater level modeling. *J. Hydrol.* **2019**, *572*, 336–351. [CrossRef]
129. Wee, W.J.; Zaini, N.A.B.; Ahmed, A.N.; El-Shafie, A. A review of models for water level forecasting based on machine learning. *Earth Sci. Inform.* **2021**, *14*, 1707–1728. [CrossRef]
130. Jones, A.; Kuehnert, J.; Fraccaro, P.; Meuriot, O.; Ishikawa, T.; Edwards, B.; Stoyanov, N.; Remy, S.L.; Weldemariam, K.; Assefa, S. AI for climate impacts: Applications in flood risk. *npj Clim. Atmos. Sci.* **2023**, *6*, 63. [CrossRef]
131. Kumar, V.; Azamathulla, H.M.; Sharma, K.V.; Mehta, D.J.; Maharaj, K.T. The state of the art in deep learning applications, challenges, and future prospects: A comprehensive review of flood forecasting and management. *Sustainability* **2023**, *15*, 10543. [CrossRef]
132. Mosavi, A.; Ozturk, P.; Chau, K.-W. Flood prediction using machine learning models: Literature review. *Water* **2018**, *10*, 1536. [CrossRef]
133. Nearing, G.; Cohen, D.; Dube, V.; Gauch, M.; Gilon, O.; Harrigan, S.; Hassidim, A.; Klotz, D.; Kratzert, F.; Metzger, A.; et al. Global prediction of extreme floods in ungauged watersheds. *Nature* **2024**, *627*, 559–563. [CrossRef] [PubMed]
134. Ng, K.W.; Huang, Y.F.; Koo, C.H.; Chong, K.L.; El-Shafie, A.; Ahmed, A.N. A review of hybrid deep learning applications for streamflow forecasting. *J. Hydrol.* **2023**, *625*, 130141. [CrossRef]
135. Karim, F.; Armin, M.A.; Ahmedt-Aristizabal, D.; Tychsen-Smith, L.; Petersson, L. A review of hydrodynamic and machine learning approaches for flood inundation modeling. *Water* **2023**, *15*, 566. [CrossRef]
136. Fleming, S.W.; Garen, D.C. Simplified cross-validation in principal component regression (PCR) and PCR-like machine learning for water supply forecasting. *JAWRA J. Am. Water Resour. Assoc.* **2022**, *58*, 517–524. [CrossRef]
137. Prodhan, F.A.; Zhang, J.; Hasan, S.S.; Sharma, T.P.P.; Mohana, H.P. A review of machine learning methods for drought hazard monitoring and forecasting: Current research trends, challenges, and future research directions. *Environ. Model. Softw.* **2022**, *149*, 105327. [CrossRef]
138. Chen, Y.; Song, L.; Liu, Y.; Yang, L.; Li, D. A review of the artificial neural network models for water quality prediction. *Appl. Sci.* **2020**, *10*, 5776. [CrossRef]
139. Zhi, W.; Appling, A.P.; Golden, H.E.; Podgorski, J.; Li, L. Deep learning for water quality. *Nat. Water* **2024**, *2*, 228–241. [CrossRef]
140. Varadharajan, C.; Appling, A.P.; Arora, B.; Christianson, D.S.; Hendrix, V.C.; Kumar, V.; Lima, A.R.; Müller, J.; Oliver, S.; Ombadi, M.; et al. Can machine learning accelerate process understanding and decision-relevant predictions of river water quality? *Hydrol. Process.* **2022**, *36*, e14565. [CrossRef]
141. Feigl, M.; Lebiedzinski, K.; Herrnegger, M.; Schulz, K. Machine-learning methods for stream water temperature prediction. *Hydrol. Earth Syst. Sci.* **2021**, *25*, 2951–2977. [CrossRef]
142. Nguyen, X.C.; Bui, V.K.H.; Cho, K.H.; Hur, J. Practical application of machine learning for organic matter and harmful algal blooms in freshwater systems: A review. *Crit. Rev. Environ. Sci. Technol.* **2023**, *54*, 953–975. [CrossRef]
143. Park, J.; Lee, W.H.; Kim, K.T.; Park, C.Y.; Lee, S.; Heo, T.Y. Interpretation of ensemble learning to predict water quality using explainable artificial intelligence. *Sci. Total Environ.* **2022**, *832*, 155070. [CrossRef] [PubMed]
144. Chen, L.; Han, B.; Wang, X.; Zhao, J.; Yang, W.; Yang, Z. Machine learning methods in weather and climate applications: A survey. *Appl. Sci.* **2023**, *13*, 12019. [CrossRef]
145. de Burgh-Day, C.O.; Leeuwenburg, T. Machine learning for numerical weather and climate modelling: A review. *Geosci. Model Dev.* **2023**, *16*, 6433–6477. [CrossRef]
146. Kaack, L.H.; Donti, P.L.; Strubell, E.; Kamiya, G.; Creutzig, F.; Rolnick, D. Aligning artificial intelligence with climate change mitigation. *Nat. Clim. Chang.* **2022**, *12*, 518–527. [CrossRef]
147. Rolnick, D.; Donti, P.L.; Kaack, L.H.; Kochanski, K.; Lacoste, A.; Sankaran, K.; Ross, A.S.; Milojevic-Dupont, N.; Jaques, N.; Waldman-Brown, A.; et al. Tackling climate change with machine learning. *ACM Comput. Surv. (CSUR)* **2022**, *55*, 1–96. [CrossRef]
148. Slater, L.J.; Arnal, L.; Boucher, M.-A.; Chang, A.Y.-Y.; Moulds, S.; Murphy, C.; Nearing, G.; Shalev, G.; Shen, C.; Speight, L.; et al. Hybrid forecasting: Blending climate predictions with AI models. *Hydrol. Earth Syst. Sci.* **2023**, *27*, 1865–1889. [CrossRef]
149. Cui, S.; Gao, Y.; Huang, Y.; Shen, L.; Zhao, Q.; Pan, Y.; Zhuang, S. Advances and applications of machine learning and deep learning in environmental ecology and health. *Environ. Pollut.* **2023**, *335*, 122358. [CrossRef]
150. Lei, L.; Pang, R.; Han, Z.; Wu, D.; Xie, B.; Su, Y. Current applications and future impact of machine learning in emerging contaminants: A review. *Crit. Rev. Environ. Sci. Technol.* **2023**, *53*, 1817–1835. [CrossRef]
151. Zhu, M.; Wang, J.; Yang, X.; Zhang, Y.; Zhang, L.; Ren, H.; Wu, B.; Ye, L. A review of the application of machine learning in water quality evaluation. *Eco-Environ. Health* **2022**, *1*, 107–116. [CrossRef]
152. Wicaksono, P.; Aryaguna, P.A.; Lazuardi, W. Benthic habitat mapping model and cross validation using machine-learning classification algorithms. *Remote Sens.* **2019**, *11*, 1279. [CrossRef]

153. Pichler, M.; Boreux, V.; Klein, A.M.; Schleuning, M.; Hartig, F. Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. *Methods Ecol. Evol.* **2020**, *11*, 281–293. [CrossRef]
154. Mosebo Fernandes, A.C.; Quintero Gonzalez, R.; Lenihan-Clarke, M.A.; Leslie Trotter, E.F.; Jokar Arsanjani, J. Machine learning for conservation planning in a changing climate. *Sustainability* **2020**, *12*, 7657. [CrossRef]
155. Tuia, D.; Kellenberger, B.; Beery, S.; Costelloe, B.R.; Zuffi, S.; Risse, B.; Mathis, A.; Mathis, M.W.; Van Langevelde, F.; Burghardt, T.; et al. Perspectives in machine learning for wildlife conservation. *Nat. Commun.* **2022**, *13*, 792. [CrossRef]
156. Anjum, R.; Parvin, F.; Ali, S.A. Machine Learning Applications in Sustainable Water Resource Management: A Systematic Review. In *Emerging Technologies for Water Supply, Conservation and Management*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 29–47.
157. Pollard, J.A.; Spencer, T.; Jude, S. Big Data Approaches for coastal flood risk assessment and emergency response. *Wiley Interdiscip. Rev. Clim. Change* **2018**, *9*, e543. [CrossRef]
158. Sun, A.Y.; Scanlon, B.R. How can Big Data and machine learning benefit environment and water management: A survey of methods, applications, and future directions. *Environ. Res. Lett.* **2019**, *14*, 073001. [CrossRef]
159. Tahmasebi, P.; Kamrava, S.; Bai, T.; Sahimi, M. Machine learning in geo- and environmental sciences: From small to large scale. *Adv. Water Resour.* **2020**, *142*, 103619. [CrossRef]
160. Zheng, H.; Liu, Y.; Wan, W.; Zhao, J.; Xie, G. Large-scale prediction of stream water quality using an interpretable deep learning approach. *J. Environ. Manag.* **2023**, *331*, 117309. [CrossRef]
161. Shen, C. A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resour. Res.* **2018**, *54*, 8558–8593. [CrossRef]
162. Zhong, S.; Zhang, K.; Bagheri, M.; Burken, J.G.; Gu, A.; Li, B.; Ma, X.; Marrone, B.L.; Ren, Z.J.; Schrier, J.; et al. Machine learning: New ideas and tools in environmental science and engineering. *Environ. Sci. Technol.* **2021**, *55*, 12741–12754. [CrossRef]
163. Muñoz-Carpena, R.; Carmona-Cabrero, A.; Yu, Z.; Fox, G.; Batelaan, O. Convergence of mechanistic modeling and artificial intelligence in hydrologic science and engineering. *PLoS Water* **2023**, *2*, e0000059. [CrossRef]
164. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; Prabhat, F. Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, *566*, 195–204. [CrossRef]
165. Sandhu, N.; Finch, R. Application of Artificial Neural Networks to the Sacramento-San Joaquin Delta. In Proceedings of the Estuarine and Coastal Modeling, San Diego, CA, USA, 26–28 October 1995; ASCE: Reston, VA, USA, 1995; pp. 490–504.
166. California Department of Water Resources. DSM2 Water Temperature Modeling Input Extension: 1922–2015. In *Methodology for Flow and Salinity Estimates in the Sacramento-San Joaquin Delta and Suisun Marsh: 43rd Annual Progress Report*; California Department of Water Resources: Sacramento, CA, USA, 2022.
167. California Department of Water Resources. Disinfection By-product Formation. In *Methodology for Flow and Salinity Estimates in the Sacramento-San Joaquin Delta and Suisun Marsh: 16th Annual Progress Report*; California Department of Water Resources: Sacramento, CA, USA, 1995.
168. California Department of Water Resources. Carriage Water. In *Methodology for Flow and Salinity Estimates in the Sacramento-San Joaquin Delta and Suisun Marsh: 16th Annual Progress Report*; California Department of Water Resources: Sacramento, CA, USA, 1995.
169. California Department of Water Resources. Emulation of DWRDSM Using Artificial Neural Networks and Estimation of Sacramento River Flow. In *Methodology for Flow and Salinity Estimates in the Sacramento-San Joaquin Delta and Suisun Marsh: 17th Annual Progress Report*; California Department of Water Resources: Sacramento, CA, USA, 1996.
170. Senevirante, S. Short-Term Improvements to Artificial Neural Network Implementation in CALSIM. In *Methodology for Flow and Salinity Estimates in the Sacramento-San Joaquin Delta and Suisun Marsh: 23rd Annual Progress Report*; California Department of Water Resources: Sacramento, CA, USA, 2002.
171. California Department of Water Resources. Artificial Neural Networks. In *Methodology for Flow and Salinity Estimates in the Sacramento-San Joaquin Delta and Suisun Marsh: 21st Annual Progress Report*; California Department of Water Resources: Sacramento, CA, USA, 2000.
172. Senevirante, S.; Wu, S. Enhanced Development of Flow-Salinity Relationships in the Delta Using Artificial Neural Networks: Incorporating Tidal Influence. In *Methodology for Flow and Salinity Estimates in the Sacramento-San Joaquin Delta and Suisun Marsh: 28th Annual Progress Report*; California Department of Water Resources: Sacramento, CA, USA, 2007.
173. Wilbur, R.; Munevar, R. Integration of CALSIM and Artificial Neural Networks Models for Sacramento-San Joaquin Delta Flow-Salinity Relationships. In *Methodology for Flow and Salinity Estimates in the Sacramento-San Joaquin Delta and Suisun Marsh: 22nd Annual Progress Report*; California Department of Water Resources: Sacramento, CA, USA, 2001.
174. Mierzwa, M. CALSIM versus DSM2 ANN and G-model Comparisons. In *Methodology for Flow and Salinity Estimates in the Sacramento-San Joaquin Delta and Suisun Marsh: 23rd Annual Progress Report*; California Department of Water Resources: Sacramento, CA, USA, 2002.

175. California Department of Water Resources. Marginal Export Cost and MDO Replacement. In *Methodology for Flow and Salinity Estimates in the Sacramento-San Joaquin Delta and Suisun Marsh: 18th Annual Progress Report*; California Department of Water Resources: Sacramento, CA, USA, 1997.
176. California Department of Water Resources. Artificial Neural Networks and MEC Estimates. In *Methodology for Flow and Salinity Estimates in the Sacramento-San Joaquin Delta and Suisun Marsh: 19th Annual Progress Report*; California Department of Water Resources: Sacramento, CA, USA, 1998.
177. Hutton, P.; Senevirante, S. An Initial Assessment of Delta Carriage Water Requirements Using a New CalSim Flow-Salinity Routine. In *Methodology for Flow and Salinity Estimates in the Sacramento-San Joaquin Delta and Suisun Marsh: 22nd Annual Progress Report*; California Department of Water Resources: Sacramento, CA, USA, 2001.
178. Saha, G.; Namadi, P.; He, M.; Sandhu, P. Machine Learning-Based Harmful Algal Blooms (HABs) Modeling. In Proceedings of the Sacramento-San Joaquin Delta, Bay-Delta Science Conference 2024, Sacramento, CA, USA, 30 September–2 October 2024.
179. Guivetchi, K. *Salinity Unit Conversion Equations*; California Department of Water Resources Interoffice Memo: Sacramento, CA, USA, 1986.
180. Hutton, P.; Sinha, A.; Roy, S. Simplified Approach for Estimating Salinity Constituent Concentrations in the San Francisco Estuary & Sacramento-San Joaquin River Delta. User Guide. July 2022. Available online: https://rtdf.info/ (accessed on 1 March 2024).
181. Bishop, C.M. *Neural Networks for Pattern Recognition*; Oxford University Press: Oxford, UK, 1995.
182. Li, B.; Tang, X.; Qi, X.; Chen, Y.; Li, C.-G.; Xiao, R. EMU: Effective multi-hot encoding net for lightweight scene text recognition with a large character set. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 5374–5385. [CrossRef]
183. Carbonell, J.G.; Michalski, R.S.; Mitchell, T.M. An overview of machine learning. *Mach. Learn.* **1983**, 3–23. [CrossRef]
184. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
185. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
186. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
187. Shen, C.; Appling, A.P.; Gentine, P.; Bandai, T.; Gupta, H.; Tartakovsky, A.; Baity-Jesi, M.; Fenicia, F.; Kifer, D.; Li, L.; et al. Differentiable modelling to unify machine learning and physical models for geosciences. *Nat. Rev. Earth Environ.* **2023**, *4*, 552–567. [CrossRef]
188. Feng, D.; Beck, H.; de Bruijn, J.; Sahu, R.K.; Satoh, Y.; Wada, Y.; Liu, J.; Pan, M.; Lawson, K.; Shen, C. Deep dive into hydrologic simulations at global scale: Harnessing the power of deep learning and physics-informed differentiable models (δHBV-globe1. 0-hydroDL). *Geosci. Model Dev.* **2024**, *17*, 7181–7198. [CrossRef]
189. Song, Y.; Knoben, W.J.; Clark, M.P.; Feng, D.; Lawson, K.; Sawadekar, K.; Shen, C. When ancient numerical demons meet physics-informed machine learning: Adjoint-based gradients for implicit differentiable modeling. *Hydrol. Earth Syst. Sci.* **2024**, *28*, 3051–3077. [CrossRef]
190. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
191. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
192. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
193. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [CrossRef]
194. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
195. Umirbekov, A.; Peña-Guerrero, M.D.; Didovets, I.; Apel, H.; Gafurov, A.; Müller, D. The Value of Hydroclimatic Teleconnections for Snow-based Seasonal Streamflow Forecasting. *Hydrol. Earth Syst. Sci. Discuss.* **2024**, 1–19. [CrossRef]
196. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B (Methodol.)* **1974**, *36*, 111–133. [CrossRef]
197. Bergstra, J.; Bardenet, R.; Bengio, Y.; Kégl, B. Algorithms for hyper-parameter optimization. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Granada, Spain, 12–15 December 2011; pp. 2546–2554.
198. Bergstra, J.; Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
199. Bergstra, J.; Yamins, D.; Cox, D. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In Proceedings of the International Conference on Machine Learning PMLR, Atlanta, GA, USA, 16–21 June 2013; pp. 115–123.
200. Baydin, A.G.; Pearlmutter, B.A.; Radul, A.A.; Siskind, J.M. Automatic differentiation in machine learning: A survey. *J. Mach. Learn. Res.* **2018**, *18*, 1–43.
201. Beyer, H.G.; Schwefel, H.P. Evolution strategies—A comprehensive introduction. *Nat. Comput.* **2002**, *1*, 3–52. [CrossRef]
202. Bischl, B.; Binder, M.; Lang, M.; Pielok, T.; Richter, J.; Coors, S.; Thomas, J.; Ullmann, T.; Becker, M.; Boulesteix, A.L.; et al. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2023**, *13*, e1484. [CrossRef]

203. Jones, D.R.; Schonlau, M.; Welch, W.J. Efficient global optimization of expensive black-box functions. *J. Glob. Optim.* **1998**, *13*, 455–492. [CrossRef]
204. Li, L.; Jamieson, K.; DeSalvo, G.; Rostamizadeh, A.; Talwalkar, A. Hyperband: A novel bandit-based approach to hyperparameter optimization. *J. Mach. Learn. Res.* **2018**, *18*, 1–52.
205. Lorraine, J.; Vicol, P.; Duvenaud, D. Optimizing millions of hyperparameters by implicit differentiation. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Toronto, ON, Canada, 26–28 August 2020; pp. 1540–1552.
206. Loh, W.Y. Classification and regression trees. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2011**, *1*, 14–23. [CrossRef]
207. Dwivedi, R.; Dave, D.; Naik, H.; Singhal, S.; Omer, R.; Patel, P.; Qian, B.; Wen, Z.; Shah, T.; Morgan, G.; et al. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Comput. Surv.* **2023**, *55*, 1–33. [CrossRef]
208. Grieves, M.W. Product lifecycle management: The new paradigm for enterprises. *Int. J. Prod. Dev.* **2005**, *2*, 71–84. [CrossRef]
209. Grieves, M.W. Digital twins: Past, present, and future. In *The Digital Twin*; Springer International Publishing: Cham, Switzerland, 2023; pp. 97–121.
210. Cannon, A.J.; McKendry, I.G. A graphical sensitivity analysis for statistical climate models: Application to Indian monsoon rainfall prediction by artificial neural networks and multiple linear regression models. *Int. J. Clim.* **2002**, *22*, 1687–1708. [CrossRef]
211. Karpatne, A.; Atluri, G.; Faghmous, J.H.; Steinbach, M.; Banerjee, A.; Ganguly, A.; Shekhar, S.; Samatova, N.; Kumar, V. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 2318–2331. [CrossRef]
212. Raissi, M.; Perdikaris, P.; Karniadakis, G.E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **2019**, *378*, 686–707. [CrossRef]
213. Cuomo, S.; Di Cola, V.S.; Giampaolo, F.; Rozza, G.; Raissi, M.; Piccialli, F. Scientific machine learning through physics–informed neural networks: Where we are and what's next. *J. Sci. Comput.* **2022**, *92*, 88. [CrossRef]
214. He, Q.; Tartakovsky, A.M. Physics-informed neural network method for forward and backward advection-dispersion equations. *Water Resour. Res.* **2021**, *57*, e2020WR029479. [CrossRef]
215. Yu, J.; Lu, L.; Meng, X.; Karniadakis, G.E. Gradient-enhanced physics-informed neural networks for forward and inverse PDE problems. *Comput. Methods Appl. Mech. Eng.* **2022**, *393*, 114823. [CrossRef]
216. Li, X.; Feng, M.; Ran, Y.; Su, Y.; Liu, F.; Huang, C.; Shen, H.; Xiao, Q.; Su, J.; Yuan, S.; et al. Big Data in Earth system science and progress towards a digital twin. *Nat. Rev. Earth Environ.* **2023**, *4*, 319–332. [CrossRef]
217. Brocca, L.; Barbetta, S.; Camici, S.; Ciabatta, L.; Dari, J.; Filippucci, P.; Massari, C.; Modanesi, S.; Tarpanelli, A.; Bonaccorsi, B.; et al. A Digital Twin of the terrestrial water cycle: A glimpse into the future through high-resolution Earth observations. *Front. Sci.* **2024**, *1*, 1190191. [CrossRef]
218. Foroumandi, E.; Moradkhani, H.; Sanchez-Vila, X.; Singha, K.; Castelletti, A.; Destouni, G. ChatGPT in hydrology and earth sciences: Opportunities, prospects, and concerns. *Water Resour. Res.* **2023**, *59*, e2023WR036288. [CrossRef]
219. Stokel-Walker, C.; Nordeen, V.R. The promise and peril of generative AI. *Nature* **2023**, *614*, 214–216. [CrossRef]
220. Zhu, J.J.; Jiang, J.; Yang, M.; Ren, Z.J. ChatGPT and environmental research. *Environ. Sci. Technol.* **2023**, *57*, 17667–17670. [CrossRef]
221. Feynman, R.P. Simulating Physics with Computers. *Int. J. Theor. Phys.* **1982**, *21*, 467–488. [CrossRef]
222. Shor, P.W. Algorithms for quantum computation: Discrete logarithms and factoring. In Proceedings of the 35th Annual Symposium on Foundations of Computer Science, Santa Fe, NM, USA, 20–22 November 1994; pp. 124–134.
223. Google Quantum AI and Collaborators. Quantum error correction below the surface code threshold. *Nature* **2024**, *638*, 920–926. [CrossRef]
224. Saygin, A.P.; Cicekli, I.; Akman, V. Turing test: 50 years later. *Minds Mach.* **2000**, *10*, 463–518. [CrossRef]
225. Barr, A.; Feigenbaum, E.A.; Cohen, P.R. (Eds.) *The Handbook of Artificial Intelligence*; HeurisTech Press: Los Altos, CA, USA, 1981; Volume 1.
226. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef] [PubMed]
227. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
228. Razavi, S. Deep learning explained: Fundamentals, explainability, and bridge ability to process-based modelling. *Environ. Model. Softw.* **2021**, *144*, 105159. [CrossRef]
229. Hassoun, M.H. *Fundamentals of Artificial Neural Networks*; MIT Press: Cambridge, MA, USA, 1995.
230. LeCun, Y.; Boser, B.; Denker, J.; Henderson, D.; Howard, R.; Hubbard, W.; Jackel, L. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1989; pp. 396–404.
231. Elman, J.L. Finding structure in time. *Cogn. Sci.* **1990**, *14*, 179–211. [CrossRef]

232. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
233. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.