Comparing machine learning models and physics-based models in groundwater science

by

Thomas Christiaan Boerman

BSc., Utrecht University, 2011

MSc., Utrecht University, 2013

A thesis submitted in partial fulfillment for the degree of

MASTER OF APPLIED SCIENCE

in the Department of Civil Engineering

© Thomas Christiaan Boerman, 2022

University of Victoria

All rights reserved. This thesis may not be reproduced in whole or in part, by photocopy or other means, without the permission of the author

Comparing machine learning models and physics-based models in groundwater science

by
Thomas Christiaan Boerman
BSc., Utrecht University, 2011
MSc., Utrecht University, 2013

Supervisory committee

Dr. Tom Gleeson (Department of Civil Engineering)
Supervisor

Dr. Chris Kennedy (Department of Civil Engineering)
Department member

Dr. Ralph Evins (Department of Civil Engineering)
Department member

Dr. Daniel Lee Peters (Department of Geography)
Outside member

Thesis abstract

The use of machine learning techniques in tackling hydrological problems has significantly increased over the last decade. Machine learning tools can provide alternatives or surrogates to complex and comprehensive methodologies such as physics-based numerical models. Machine learning algorithms have been used in hydrology for estimating streamflow, runoff, water table fluctuations and calculating the impacts of climate change on nutrient loading among many other applications. In recent years we have also seen arguments for and advances in combining physics-based models and machine learning algorithms for mutual benefit. This thesis contributes to these advances by addressing two different groundwater problems by developing a machine learning approach and comparing this previously developed physics-based models: i) estimating groundwater and surface water depletion caused by groundwater pumping using artificial neural networks and ii) estimating a global steady-state map of water table depth using random forests.

The first chapter of this thesis outlines the purpose of this thesis and how this thesis is a contribution to the overall scientific knowledge on the topic. The results of this research contribute to three of the twenty-three major unsolved problems in hydrology, as has been summarized by a collective of hundreds of hydrologists.

In the second chapter, we tested the potential of artificial neural networks (ANNs), a deep-learning tool, as an alternative method for estimating source water of groundwater abstraction compared to conventional methods (analytical solutions and numerical models). Surrogate ANN models of three previously calibrated numerical groundwater models were developed using hydrologically meaningful input parameters (e.g., well-stream distance and hydraulic diffusivity) selected by predictor parameter optimization, combining hydrological expertise and statistical methodologies (ANCOVA). The output parameters were three transient sources of groundwater abstraction (shallow and deep storage release, and local surface-water depletion). We found that the optimized ANNs have a predictive skill of up to 0.84 (R^2, $2\sigma = \pm 0.03$) when predicting water sources compared to physics-based numerical (MODFLOW) models. Optimal ANN skill was obtained when using between five and seven predictor parameters, with hydraulic diffusivity and mean aquifer thickness being the most important predictor parameters. Even though initial results are promising and computationally frugal, we found that the deep learning models were not yet sufficient or outperforming numerical model simulations.

The third chapter used random forests in mapping steady-state water table depth on a global scale (0.1°-spatial resolution) and to integrate the results to improve our understanding on scale and perceptual modeling of global water table depth. In this study we used a spatially biased ~1.5-million-point database of water table depth observations with a variety of

globally distributed above- and below-ground predictor variables with causal relationships to steady-state water table depth. We mapped water table depth globally as well as at regional to continental scales to interrogate performance, feature importance and hydrologic process across scales and regions with varying hydrogeological landscapes and climates. The global water table depth map has a correlation (cross validation error) of $R^2 = 0.72$ while our highest continental correlation map (Australia) has a correlation of $R^2 = 0.86$. The results of this study surprisingly show that above-ground variables such as surface elevation, slope, drainage density and precipitation are among the most important predictor parameters while subsurface parameters such as permeability and porosity are notably less important. This is contrary to conventional thought among hydrogeologists, who would assume that sub-surface parameters are very important. Machine learning results overall underestimate water table depth similar to existing global physics-based groundwater models which also have comparable differences between existing physics-based groundwater models themselves. The feature importance derived from our random forest models was used to develop alternative perceptual models that highlight different water table depth controls between areas with low relief and high relief. Finally, we considered the representativeness of the prediction domain and the predictor database and found that 90% of the prediction domain has a dissimilarity index lower than 0.75. We conclude that we see good extrapolation potential for our random forest models to regions with unknown water table depth, except for some high elevation regions.

Finally in chapter four, the most important findings of chapters two and three are considered as contributions to the unresolved questions in hydrology. Overall, this thesis has contributed to advancing hydrological sciences through: i) mapping of global steady-state water table depth using machine learning; ii) advancing hybrid modeling by using synthetic data derived from physics-based models to train an artificial neural network for estimating storage depletion; and (iii) it contributing to answering three unsolved problems in hydrology involving themes of parameter scaling across temporal and spatial scales, extracting hydrological insight from data, the use of innovative modeling techniques to estimate hydrological fluxes/states and extrapolation of models to no-data regions.

Table of contents

List of Tables

List of Figures

of the base model (Table 2.4 and Figure 2.5). Every independent data point represents an artificial neural network where one predictor is omitted. The annotations depict the predictors that were dropped at each subsequent step, since it was the least contributing or redundant predictor parameter (as is explained in section (2.2.6). The colors represent the range in R^2_{NN}over 2 standard deviations over 10-fold cross validation.

Figure 2.7 Correlations of relative contribution of response variable to the total abstraction rate ($\frac{Q_{source}}{Q_{well}}$) and database skewness against neural network still (R^2_t). The three colors represent the three response variables.

Figure 2.8 Heat map showing the ANN predictor skill $(R^2_{ANN})$for cross-basin comparison averaged over 10-fold cross validation. Training datasets are given on the vertical axis and testing data is shown on the horizontal axis. Subplots a), b) and c) represent scatter plots for an ANN trained on the KAMA database and test data used from the a) KALA, b) KAMA and c) KAMAUP database. The samples are marked by database origin (KALA: green, MANI: blue and UPFOX: red).

Figure 3.1 Schematic representations of hydrogeological studies involving a) developing numerical models and b) developing machine learning models. Solid arrows indicate steps in the modeling process, while dashed arrows depict feedback loops on how the results are beneficial for gaining scientific knowledge. Icons are obtained from the Noun Project (*Noun Project: Free Icons & Stock Photos for Everything*, n.d.) and Wagener et al. (2021)

Figure 3.2 Data density of water table depth measurements within the Fan et al database, resampled to 0.1° spatial resolution (Fan et al., 2013).

Figure 3.3 Workflow, references to tables and figures and deliverables of this study indicating the three main contributions to the scientific understanding of machine learning models predicting water table depth.

Figure 3.4 Example of the building of the final water table depth maps in this study as the sum of areas of known data (training domain (top left)) and the regions of unknown water table depth where we extrapolated our model to (predicted domain (top right)). The combination gives the final water table depth map (bottom). The full maps for all regions are found in the Supplementary information.

Figure 3.5 Random forest model results for water table depth for the Global model (top) and every sub database: a) United States (excluding Alaska, Hawaii and overseas territories), b) Iberia, c) France, d) Rhine Delta, e) Colorado, f) Brazil, g) Australia and h) California. The scatter plots show the cross-validation errors and correlations (mean absolute error (MAE), root

mean squared error (RMSE) and correlation (R^2) for every model. The diagonal represents the 1:1 correlation line.

Figure 3.6 Results of feature importance analysis. Left: Ranked feature importance as calculated by Scitkit-learn's feature importance function. The horizontal bars show the feature importance within the Global database. The error bars show the standard deviation of each individual predictor over every other database. Right: reduction in feature importance as predictor is replaced by randomized data.

Figure 3.7 Global map of water table depth (m) based on the random forest model (for the predicted domain) combined with the training data from the Fan et al database on a 0.1-degree spatial resolution.

Figure 3.8 Pair grid plot of model correlation between four distinct models: 1) The random forest model for water table depth (this study), 2) Fan et al model (Fan et al., 2013), 3) G^3M model (Reinecke et al., 2019) and 4) De Graaf model (De Graaf et al., 2015). The figures show heatmaps of water table depth measurements (darker shade indicates higher point density) and the solid lines depicts 1:1 lines for WTD.

Figure 3.9 Representation of this underlying study based on the schematic representation of geosciences within a machine learning framework (Figure 1b). Subfigures: a) the feature importance for the Global database (green) and the feature importance for the Rhine Delta database. b) Translation of the feature importance from to two conceptual models based on the results of the random forest model for an area with low relief (orange) and for an area with high and low relief (green).

Figure 3.10 Importance of scale when predicting water table depth using random forests based on different training datasets (y-axis) and prediction domains (x-axis). Correlation values between maps are given in boxes. Example of correlation plot is given for two maps.

Figure 3.11 a) Dissimilarity Index calculated for the predicted domain on a 0.1° scale. Methodology based on the studies by Meyer and Pebesma (2020). b) Scatter density plot of predicted water table depth by the Global model against DI. Colors show point density. Outlier values for DI (>2) are excluded from this graph. c) Cumulative Distribution Function (CDF) of global distribution of DI.

Figure B-1 Example of kernel distributions for predictor parameter DD within the Fan et al database (DB, blue) and the global distribution (Global). The overlap index (OI) is calculated as the percentage overlap between the areas under the kernel density distribution curves. The three graphs (from left to right) show the results for 0.1°, 0.25° and 0.5° spatial resolution.

Author Contributions

The core of this manuscript-based thesis consists of two chapters that will be submitted to scientific journals.

Chapter 2 is a manuscript titled: "*Predicting the transient sources of groundwater abstraction using Artificial Neural networks: possibilities and limitations"* with the following authors:

Boerman, Thomas C.[1], Gleeson, T.[1], Zipper, S.C.[1], Li, Q. [1], Greve, P.[2], Wada, Y.[2]
[1] Department of Civil Engineering, University of Victoria, BC, Canada
[2] International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria

TB developed the methodology, built the models and is the lead author on the manuscript. TG supervised and contributed to the methodology, interpretation, and presentation of the results. PG and YW supervised TB during this semester at the International Institute of Applied Systems Analysis (IIASA) in Laxenburg, Austria and contributed to the methodology, interpretation, and presentation of the results. SZ contributed to the methodology, interpretation, and presentation of the results.

Chapter 3 is a manuscript titled: "*Disentangling process controls on global groundwater table depth patterns using random forests*" with the following authors:

Boerman[1], Thomas C., Gleeson[1], T., Reinecke[2], R., Wagener, T.[2]
[1] Department of Civil Engineering, University of Victoria, BC, Canada
[3] Institute of Environmental Science and Geography, University Potsdam, Germany

TB developed the methodology, built the models and is the lead author on the manuscript. TG supervised and contributed to the methodology, interpretation, and presentation of the results. RR and TW contributed to the methodology, interpretation, and presentation of the results.

Acknowledgements

This research was funded by the Canadian Natural Sciences and Engineering Research Council (NSERC). Part of the research was developed in the Young Scientists Summer Program at the International Institute for Systems Analysis, Laxenburg (Austria) with financial support from the Netherlands Organisation for Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO).

The author has no conflict of interest regarding the results and conclusions of this research which would prohibit its publication or release.

All source data used in this study, including numerical groundwater models are open source and available for download through the United States Geological Survey's website or through contacting the corresponding authors in the cited studies. Results, tables, figures, models and/or Python code will be made available through Github or can be provided by the author of this thesis upon request.

Personal thanks

I want to thank so many people who have helped me throughout the years. Without their help I would never have made it through grad school.

First and foremost, I want to thank my supervisor Tom Gleeson for inspiring me, helping me improve my skills and continuously believing in me. I can safely say that I learned from you more than I have learned from any teacher in my life.

Next, I want to thank my committee and all my fellow students and coworkers in the Groundwater Science and Sustainability Group: Tara, Mikhail, Xander, Leigh, Tom, Jordan, Sam, John, Ashley, Chinchu, David, Kristina, and Sacha. It was a blast sharing an office with you all. Many funny group meetings, discussions, and coffee breaks.

I also want to give a shout out to my other UVic grad students, which I spend multiple nights with "singing" karaoke or hanging out at the board game cafe.

Last and certainly not least, I want to thank my friends and family in the Netherlands for the warm welcomes when I came home for Christmas break. And above all my thanks go out to my brother Marcel, sister-in-law Heleen, cousin Wout and my parents Ernst and Annerie Boerman who have always continued to support and believe in me, even during times when I thought I would never finish this program. I couldn't have done it without all your support.

Lots of love,
Thomas

Personal quote

"In a dark place we find ourselves. And a little more knowledge lights our way."
—Yoda, *Revenge of the Sith (2005)*

Chapter 1: Purpose and Thesis Contributions

This chapter provides a general introduction to machine learning and its application in hydrology to contextualize the purpose and contributions of this thesis. More detailed and focused literature review is included in Chapters 2 and 3.

1.1 Introduction to machine learning

The earliest mention of the term "machine learning" in the scientific literature was found in the 1930's, but it was not until the 1980's that the number of studies involving machine learning started to grow exponentially (Thessen, 2016). Nowadays, there are few fields within the entire science community where machine learning algorithms have not been used. Within popular culture, machine learning evokes thoughts of robotics or science-fiction, but machine learning has very useful, basic day-to-day applications and uses. One of the primary advantages of machine learning is its transferability among different fields of science. Though different fields use different metrics, models, or scales to analyze problems, the mathematical methodology of machine learning is transferable across nearly any scales, domains, and settings. As can be seen in Figure 1.1, machine learning is a nested subsection of artificial intelligence. Machine learning can also be called an umbrella term for a large variety of different algorithms, each with their own strengths, weaknesses and uses.

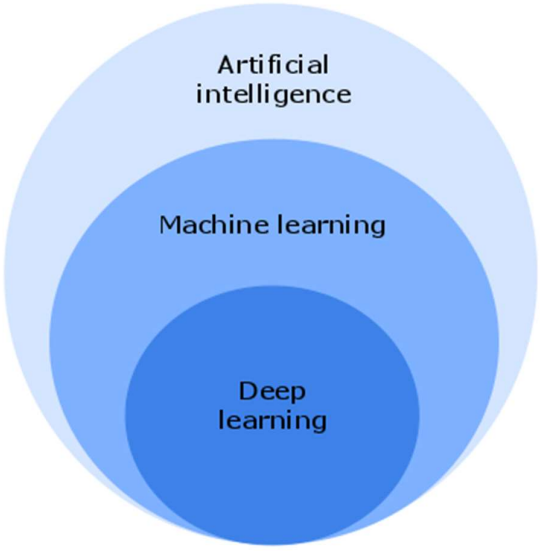

Figure 1.1 Schematic representation of major nomenclature after Mitchell (1997) and Xu & Liang (2021)

Machine learning has been introduced in many branches of science:

- In the physical sciences we have seen it being used in particle physics, cosmology, or quantum computing among numerous other applications (Carleo et al., 2019).
- In the social sciences we have seen studies involving crime prediction using machine learning algorithms (Reier Forradellas et al., 2021) and in the last few years we have

seen several articles promoting the benefits of machine learning algorithms to the social sciences (Grimmer et al., 2021; Hindman, 2015).

- Within the field of biology multiple machine learning algorithms have been compared to explain the declining populations of ocellated turkeys (Kampichler et al., 2010) and in structural biology to predict the secondary structures of proteins (Sternberg et al., 1994).
- In the medical sciences machine learning has been used to predict breast cancer survival (Montazeri et al., 2016).

This is an impressive range of utilities for a set of algorithms and therefore has proven its use extensively. However, most people probably know of machine learning and its use in popular social media applications and app features such as facial recognition, voice detecting or fingerprint scanning. All these features use machine learning algorithms to some degree.

Machine learning has also made its way into the hydrological sciences (Nearing et al., 2021). Currently, machine learning applications in hydrology are booming as evidenced by recent review and perspective papers (Reichstein et al., 2019; Shen, 2018; Shen et al., 2021; Shen & Lawson, 2021; Xu & Liang, 2021). This rapid rise of the use of machine learning algorithms in hydrology is an exciting and promising development for hydrologists and anyone else with an interest in water. This is supported by a recent special edition of the journal of *Water* that gave 14 examples of studies involving Artificial Intelligence advancing hydrologic forecasts and water resource management (Chang & Guo, 2020). This Master's research contributes to this evolution of machine learning in new and innovative ways. To motivate and contextualize these contributions within the hydrological and geological science literature, this chapter includes a (limited) literature review of 1) the most recent advances of machine learning in hydrology and 2) the dominant explored and unexplored problems within the field of hydrology and on how machine learning could be used to further our hydrologic understanding.

1.2 Recent developments on the use of machine learning in hydrology

Linear or multivariate regression models are among the simplest forms of machine learning where a dataset of one or more independent variables are approximated by a polynomial function of one or more dependent variables. These relatively simple models are quick and easy approximations to complex problems and were suggested early on for their use within hydrology (DeCoursey & Deal, 1974; Snyder, 1962). The simplest forms are used less often in the last few years, but an advanced version of this called multivariate adaptive regression splines was used to predict streamflow and land use change for example (Adnan et al., 2020; Al-Sudani et al., 2019). Some of the other popular and more advanced machine learning algorithms involve support vector machines, artificial neural networks, Bayesian networks, clustering, and random forests (Alpaydin, 2014). Each of these algorithms have been widely used to solve various hydrological problems. For example, support vector machines have

been used to estimate downscaled versions of climate parameters and to estimate streamflow (Adnan et al., 2017; Deka, 2014; Tripathi et al., 2006). Artificial neural networks have been used to simulate runoff and to simulate groundwater level fluctuations (Daliakopoulos et al., 2005; Tokar & Johnson, 1999). K-means clustering has been used to divide the world in separate hydrogeologic regions (Reinecke et al., 2020). Bayesian networks have been used to predict the impacts of climate change on nutrient loading (Sperotto et al., 2019). Finally, random forests are versatile in that they can be used for both classification problems, as well as regression problems (Brown et al., 2014; Liaw & Wiener, 2002). An example of a classification study involving random forests was done by Brown et al. where they used random forests to make a hydrologic landscape regionalization based on the hydrologic landscape concept by Winter (2001). Most of these algorithms (random forests, artificial neural networks, multivariate regression, support vector machines and Bayesian networks) are examples of *supervised learning*. In supervised learning the target variable has a known label or number and we fit our model to our known outcomes. K-means clustering is an example of *unsupervised learning*, where the target variable, number of classes is not known prior to the study and the goal of the study is to find a label or class for each sample.

One interesting and expanding part of machine learning is deep learning (Figure 1.1). Shen (Shen, 2018; Shen & Lawson, 2021) has recently done an extensive overview of deep learning research and relevance for hydrologists and water resource engineers. Although there is not a strict definition, deep learning often refers to multi-layer neural networks trained on large datasets. The term 'deep' comes from the experience that neural networks with multiple hidden layers are more likely to extract complex abstract features from raw data better than non-deep neural networks can (Goh et al., 2017; Schmidhuber, 2015). Deep learning is currently being used in hydrology: i) extract hydrometeorological and hydrologic information from images, ii) dynamically model of hydrologic variables or data collected by sensor networks and iii) learn and generate complex data distributions (Shen, 2018; Shen & Lawson, 2021). However, Shen interestingly states that we do not yet have any studies on interpreting deep learning models. The same paper lists the possible uses and risks of using deep learning within hydrology. According to Shen et al. (2021), unexplored territories of machine learning within hydrology includes vegetation hydraulics, glaciers, preferential flow, hyporheic exchange, and regional groundwater recharge among potential others. Most studies involve single output parameters that are predicted (e.g., precipitation or streamflow). Furthermore, recent studies suffered from relatively small datasets, limiting their use outside of the training domain, and making them less suitable as extrapolation tools (Reichstein et al., 2019). Multi-task models (models that predict multiple parameters simultaneously predicted) could be an innovative addition to the current scientific knowledge and practice.

Within the broader field of geoscience deep learning has become one of the preferred methods to deal with remote sensing data (L. Zhang et al., 2016). For example, remote sensing data is used for observing geometric shapes (rivers, hillslopes, valleys etc.) and convolutional

neural networks are excellent at interpreting observations from image data (Makantasis et al., 2015). Climate studies where deep learning was used involved precipitation forecasting (Hernández et al., 2016; Shi et al., 2017; P. Zhang et al., 2017) and identification of extreme weather events (Iglesias et al., 2015; Liu et al., 2016; Mudigonda et al., 2021; Prabhat et al., 2017).

In the last five years a trend has developed to combine physics-based models with machine learning models (Bhasme et al., 2021a; Willard et al., 2020; Yang et al., 2019). Common criticisms of machine learning algorithms are 1) the models are in essence a "black-box" neglecting any of the known physical laws that control groundwater flow or other hydrologic processes and 2) tend to underperform against process- or physics-based models. Kratzert (2019) states "that it is often argued that data-driven models might underperform relative to models that include explicit process representations in conditions that are dissimilar to training data", but they also argue that to their knowledge this hypothesis has not been tested to this day. There have been some attempts to combine data-driven and physics-based models and recent studies have shown that a combination or hybrid modelling setup where machine learning and conventional methods are combined can be successful. For example, Khandelwal (2020) used a physics guided approach to predict streamflow. The ensemble model in their study predicted evapotranspiration, soil water and snowpack independently using recurrent neural networks. The results were fed to another recurrent neural network which was trained using a physics-based loss function. The loss function related streamflow to precipitation, evapotranspiration, and changes in water storage to streamflow based on the principle of conservation of mass. This study is an example of how expertly chosen physical laws and machine learning algorithms could help improve each other.

One final topic that has gained popularity in recent years is inference of hydrological processes or insight from machine learning models. Marçais and de Dreuzy (2017) laid out a roadmap for the relevance of deep learning in three steps: 1) testing on data, 2) testing on benchmarks and 3) collaboration with the wider deep learning community. Within step 1, Marçais and de Dreuzy (2017) argue that both measured as well as synthetic databases could be valuable to the overall goal, since the availability of such databases has been the key to the success of deep learning studies in the past (Krizhevsky et al., 2012). There is also precedent for learning on synthetic data. Previous machine learning studies have used synthetically derived data from numerical (MODFLOW) models to generate large quantities of data for machine learning algorithms to train. These machine learning models trained on synthetic data are sometimes called *metamodels* and are functioning as a statistical model based on a computationally intensive numerical model (Feinstein et al., 2016; Fienen et al., 2015, 2016).

1.3 The unsolved problems in hydrology and the potential of machine learning

A 2019 paper published by 230 scientists within the hydrological science community outlined the twenty-three unsolved problems within hydrology (Blöschl et al., 2019; Table 1.1). From this table three problems are highlighted which lie within the scope of this thesis. These questions deal with problems in spatial variability (mainly of input parameters and processes) and how data and modeling methods can help us to gain more understanding or how to make more advanced versions of numerical models.

Table 1.1 The twenty-three unsolved problems in hydrology (according to Blöschl et al. 2019). This dissertation contributes to answering the questions highlighted in bold.

Time variability and change 1. Is the hydrological cycle regionally accelerating/decelerating under climate and environmental change, and are there tipping points (irreversible changes)? 2. How will cold region runoff and groundwater change in a warmer climate (e.g. with glacier melt and permafrost thaw)? 3. What are the mechanisms by which climate change and water use alter ephemeral rivers and groundwater in (semi-) arid regions? 4. What are the impacts of land cover change and soil disturbances on water and energy fluxes at the land surface, and on the resulting groundwater recharge?
Space variability and change 5. What causes spatial heterogeneity and homogeneity in runoff, evaporation, subsurface water and material fluxes (carbon and other nutrients, sediments), and in their sensitivity to their controls (e.g. snow fall regime, aridity, reaction coefficients)? **6. What are the hydrologic laws at the catchment scale and how do they change with scale?** 7. Why is most flow preferential across multiple scales and how does such behavior co-evolve with the critical zone? 8. Why do streams respond so quickly to precipitation inputs when storm flow is so old, and what is the transit time distribution of water in the terrestrial water cycle?
Variability of extremes 9. How do flood-rich and drought-rich periods arise, are they changing, and if so why? 10. Why are runoff extremes in some catchments more sensitive to land use/cover and geomorphic change than in others? 11. Why, how and when do rain-on-snow events produce exceptional runoff?
Interfaces in hydrology 12. What are the processes that control hillslope–riparian–stream–groundwater interactions and when do the compartments connect? 13. What are the processes controlling the fluxes of groundwater across boundaries (e.g. groundwater recharge, inter-catchment fluxes and discharge to oceans)? 14. What factors contribute to the long-term persistence of sources responsible for the degradation of water quality? 15. What are the extent, fate and impact of contaminants of emerging concern and how are microbial pathogens removed or inactivated in the subsurface?
Measurements and data **16. How can we use innovative technologies to measure surface and subsurface properties, states and fluxes at a range of spatial and temporal scales?** 17. What is the relative value of traditional hydrological observations vs soft data (qualitative observations from lay persons, data mining etc.), and under what conditions can we substitute space for time?

18. How can we extract information from available data on human and water systems in order to inform the building process of socio-hydrological models and conceptualizations?
Modeling methods **19. How can hydrological models be adapted to be able to extrapolate to changing conditions, including changing vegetation dynamics?** 20. How can we disentangle and reduce model structural/parameter/input uncertainty in hydrological prediction?
Interfaces with society 21. How can the (un)certainty in hydrological predictions be communicated to decision makers and the general public? 22. What are the synergies and tradeoffs between societal goals related to water management (e.g. water–environment–energy–food–health)? 23. What is the role of water in migration, urbanization and the dynamics of human civilizations, and what are the implications for contemporary water management?

1.4 Thesis Contributions

Based on the literature review in this chapter it can be concluded that there is room for improvement in machine learning in the hydrological sciences. This thesis attempts to fill some of the knowledge gaps of the aforementioned literature. From the list of twenty-three unsolved problems in hydrology, this work contributes to addressing three unsolved problems (using numbering from Table 1.1 for consistency):

6. What are the hydrologic laws at the catchment scale and how do they change with scale?
16. How can we use innovative technologies to measure surface and subsurface properties, states, and fluxes at a range of spatial and temporal scales?
19. How can hydrological models be adapted to be able to extrapolate to changing conditions (including changing vegetation dynamics)?

Here major contributions of this thesis are highlighted:

- While there have been global scale physics-based models for estimating groundwater (De Graaf et al., 2015; Fan et al., 2013; Reinecke et al., 2019; Sutanudjaja et al., 2018), water use (Alcamo et al., 2003) and streamflow (Li et al., 2015), we have not seen machine learning equivalents to these models at the global scale yet (Nearing et al., 2021). Therefore, this research will be among the first to build a machine learning model that attempts to produce hydrologically meaningful results on the global scale. This relates to unresolved problems 6 and 19 because it looks at how the importance of hydrological parameters varies over different spatial scales and if these parameters are transferable across watershed boundaries or across spatial scales (extrapolation question). This thesis does not touch upon the impact on changing vegetation dynamics (mentioned in question 19 specifically), but more on extrapolation to changing conditions in general.

- Within this work different machine learning methods (artificial neural networks and random forests) were used to highlight the different possible uses for machine learning algorithms. In terms of spatial and temporal variability, the scope of this master's research varied from the local (watershed) scale to the global scale and both steady state as well as transient simulations were performed. This part of the research relates to unsolved problem 16, because it attempts to predict fluxes/states (storage depletion and water table depth) using machine learning methods. The novelty (or innovative) part of this research is also that contrary to current research, we predict depletion using a multitask (multi-output) neural network (chapter 2) and we predict steady-state water table depth, contrary to transient water table depth (chapter 3).

- Chapter 2 of this thesis builds on the current advances of metamodels by predicting the sources of water from groundwater abstraction using a metamodel: artificial neural networks are trained to predict the transient changes in storage depletion and surface water depletion caused by groundwater pumping. The artificial neural networks are trained on synthetically generated data from numerical groundwater models. Three watersheds within the United States were selected as case studies. The matter of extrapolation to other regions was also investigated by taking different watersheds as training sets and predicting for other watersheds outside the training domain. This relates to unresolved problems questions 6 and 16, because we investigated how to predict storage depletion on the watershed/catchment scale, and we used hydrological models as our source of training data. This final part is an example of hybrid modeling: machine learning models trained on synthetic data from physics-based models.

- Chapter 3 of this thesis predicts global water table depth on 0.1° spatial resolution using a random forest model and compares the results to physics-based numerical models. Moreover, the same chapter also used the feature importance and results of the random forest model to conceptualize the important hydrological and climate parameters into a conceptual (or perceptual) model based on the spatial scale and geographical domain. Furthermore, the results compare the results of a sub-domain of a global model (in this case the United States and the state of California) to the results of continent and regional scale models of only the corresponding sub-domains.

- Finally, this research aims to contribute to the open science movement (Añel et al., 2019; Crochemore et al., 2020; Cudennec et al., 2020; Hall et al., 2021; Powers & Hampton, 2019; Reichman et al., 2011; Slater et al., 2019; Stagge et al., 2019; Wagener, 2020; Zipper et al., 2019). Data, code, and results will be made available upon request to make this work transparent, reproducible, and accessible to interested readers.

Chapter 2: Predicting the transient sources of groundwater abstraction using artificial neural networks: possibilities and limitations

2.1 Introduction

Groundwater is a crucial source of freshwater for human consumption, irrigation, and industrial purposes (Arnell, 1999; Bredehoeft et al., 1982; Llamas & Custodio, 2002). However, groundwater abstraction reduces water supplies stored in aquifers (storage depletion) and captures water which otherwise would have discharged from the aquifer to surface water features (surface-water depletion) (Barlow & Leake, 2012). Initially, aquifer storage is generally the dominant source of water abstracted by the well, while surface water depletion tends to become the dominant source as abstraction time increases (Bouwer & Maddock III, 1997; Kendy & Bredehoeft, 2006; Mair & Fares, 2010). When groundwater discharge to streams, lakes, and wetlands decreases to the point where environmental flow needs are not met, this can have potentially harmful effects on aquatic ecosystems (Dyson et al., 2003; Gleeson & Richter, 2018; Poff & Zimmerman, 2010).

A variety of methods have been used to estimate groundwater contribution to surface water and its depletion by wells. Analytical solutions are often used by hydrologists because they require only a few input parameters and are simple to use (Glover & Balmer, 1954; Hantush, 1965; Hunt, 1999; Zipper et al., 2018; Zlotnik, 2004; Zlotnik et al., 1999; Zlotnik & Tartakovsky, 2008). The disadvantage of these analytical solutions is found in the very specific and limited geometric and boundary conditions for which they were derived (e.g., isotropic aquifer of constant thickness with infinite extent, flat surface elevation). Recent testing however suggests they may be more widely applicable than previously assumed when combined with empirical methods to apportion depletion within a surface-water network (Zipper et al., 2018; Zipper, Gleeson, et al., 2019). Numerical groundwater models such as MODFLOW (Harbaugh, 2005; Harbaugh et al., 2000) are also commonly used as a tool for groundwater flow and simulations of groundwater abstraction. Numerical models, unlike analytical solutions, can incorporate complex geometries and boundary conditions and are therefore more versatile (Barlow & Leake, 2012; Konikow & Leake, 2014). However, numerical models require field measurements to calibrate model parameters and often require hydrogeological expertise to build, calibrate and use (Doherty, 2003; Haitjema et al., 2001; Hill, 2000; Khadri & Pande, 2016; Razavi & Gupta, 2016; Razavi & Tolson, 2013). Many regions, especially remote areas, lack the time, financial resources, and data to build a numerical model.

Given the limitations of both analytical solutions and numerical methods, statistical approaches (such as machine learning algorithms) provide a potential path to estimate the source of water to wells in complex hydrogeological environments where numerical models

do not, and are unlikely to, exist (Fienen et al., 2016; Fienen & Plant, 2015). The field of artificial intelligence (Goldberg & Holland, 1988)] has become popular in recent years with a variety of fields adopting different deep learning methods (and in particular multi-layer neural networks) into their sciences, including health, ecology and economy (Lek et al., 1996; Lek & Guégan, 1999; Trippi & Turban, 1992). An artificial neural network (ANN) (Figure 2.1) is a machine learning algorithm for finding relations in a dataset consisting of input parameters (*predictors*) and output parameters (*responses*) (Grossberg, 1988; Schmidhuber, 2015). Recent work has demonstrated that ANNs are capable of accurately predicting surface water depletion (Feinstein et al., 2016; Fienen et al., 2016), but so far only under steady-state conditions. Other benefits of these algorithms are that they can be relatively, computationally inexpensive, and more transferable than equivalent numerical models, given that numerical models are typically only created to be used on a specific region, watershed, or other fixed domain.

The application of machine learning within the field of hydrology includes short-term streamflow and rainfall-runoff modeling and forecasting, among others (Antonopoulos et al., 2016; Dehghani et al., 2014; Lange & Sippel, 2020; Moradkhani et al., 2004; Taormina & Chau, 2015; Tokar & Johnson, 1999; Zealand et al., 1999). Deep learning is a subdomain under the umbrella term machine learning. In deep learning (DL) complex machine learning models are trained to gain insight and knowledge on the statistical relations within data. Deep learning has been used in Earth science studies to simulate complex statistical correlations. Reichstein et al. (2019) summarized studies in which deep learning approaches could be or have been used compared to conventional approaches in various earth system science studies. These approaches range from river run-off predictions using a combination of convolutional neural networks with recurrent networks to transport modeling using a hybrid physical-convolutional neural network.

While machine learning has gained support within the hydrological scientific community, the overall question whether a 'black box' statistical model is as useful as a physics-based model (for example MODFLOW) in simulations physics-based simulations remains unanswered. Gaume and Gosset (2003) argued that the optimization process for choosing input parameter selection can change the 'optimal' parameter set found this way. Choosing predictor parameter sets based solely on 'hydrological expertise' may or may not result in the best performing machine learning model Therefore, parameter selection process must go through a rigorous selection process, either through multiple random selections (Zealand et al., 1999) or through cross validation (Coulibaly et al., 2000).

The goal of this study is to test whether an ANN can predict the transient sources of water to groundwater abstraction wells in a shallow, unconfined aquifer with similar accuracy compared to numerical models. We predict three different sources of water: 1) surface-water contributions (coming from streaks, lakes, and wetlands), and the 2) shallow groundwater (top aquifer) and 3) deep-stored groundwater contribution (deeper layers/aquifers). We used a predictor parameter selection approach for the input of our ANN, combining the values of

hydrological expertise, statistical relevance, and a comprehensive elimination system. The predictor parameters are all easily obtainable (field parameters) with known hydrogeological relationships to the expected amount of surface water depletion and changes in storage, e.g distance of well to the nearest stream, aquifer properties and stream density. The ANNs are trained on generated data from three numerical groundwater models. A fraction of the generated data is used for training, while the remaining data is used for testing. From there, we optimized our predictor parameter selection by finding the minimum number of required predictors while maintaining high ANN prediction skills. The final part of our study of this work attempts to use ANNs as a predictor tool for watersheds outside the training domain (extrapolation or cross-basin comparison). We chose to use ANNs as our algorithm due to the experience level on machine learning of the main authors (this study being among their first projects) and the wide use of this algorithm in previous studies (see Chapter 1).

We used the coefficient of determination (R^2) between the predicted value by the ANN and the value from a physics based numerical groundwater model (MODFLOW) as our performance criteria. In this study, we looked at the predictive skill at each individual output time ($R^2_{t_i}$) and the overall (mean) predictive skill over all output times combined (R^2).

With this work we aim to reach two goals:

1. **Gain insight into whether ANNs are able to simulate physics-based models**: in this study we not only offer a stepwise optimization approach, but we also analyze the optimized predictor parameter sets and discuss whether they are justifiable in a physics-based sense.
2. **ANNs as surrogate models:** test whether ANNs are usable as predictor tools for regions within or outside their training domain (cross basin comparison).

If the ANN method for predicting sources of water under transient conditions is successful, this will give hydrologists an additional method for estimating surface water depletion which is (1) easy to use and requires limited data and (2) potentially area independent, because the prediction is only based on simple, geographically independent parameters and could therefore be used in areas with limited resources and field data.

The goal of this paper is not to prove that ANNs are a better performing alternative to numerical groundwater models, but to test whether ANNs are able to simulate the complex physics-based dynamics of numerical models. Ideally, and as a future goal, these methods could be used to predict the effects of groundwater abstraction through time and to fill in the gaps of missing data in data records of observed surface water depletion, as has previously been done for streamflow hydrograph records (Moradkhani et al., 2004).

2.2 Methods

2.2.1 Artificial neural networks: predictors and response variables

ANNs are trained on a database of related predictors (input parameters) and responses (output parameters). A schematic representation of an ANN is given in Figure 2.1. An ANN consists of an *input layer* with one node per input parameter, an *output layer* with one node per response and several one or more layers in between, also known as hidden layers. The number of input nodes, output nodes, the number of hidden layers and the number of nodes per hidden layer are variable and dependent on the database and the scope of the problem. Neural network structure is often optimized through trial and error (Maier & Dandy, 2000).

The response variables can be either categorical or regression-based determinations, either to a single response variable or to multiple response variables simultaneously. This last form has been used in this study to predict the sources of water to wells for five different output times (after 0.5 yr, 1 yr, 5 yrs, 10 yrs and 25 yrs of abstraction, Figure 2.1). These five points together produce data points on which the trend of depletion can be estimated.

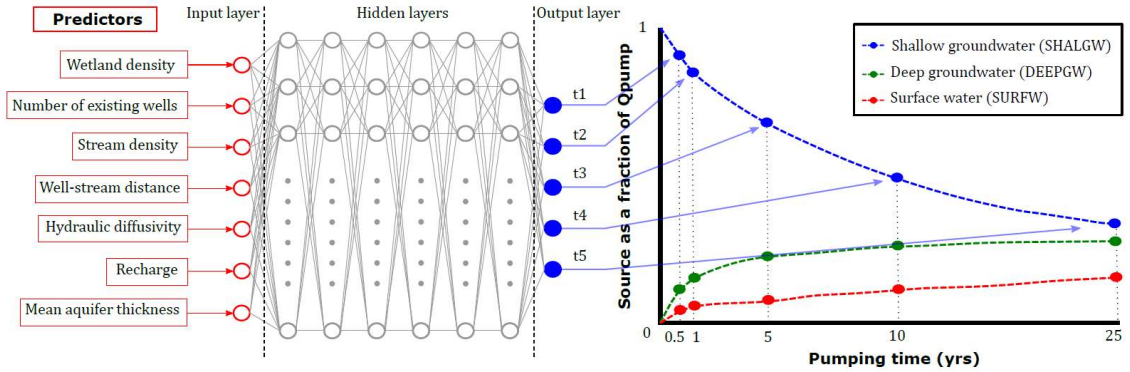

Figure 2.1 Schematic depiction of the process of using predictor parameters to predict sources of water to wells at different output times. For each source of water (shallow groundwater, deep groundwater or surface water) a different ANN is used, but each output node predicts a different output time simultaneously. The input nodes of the ANN represent the predictors used in this study

The general trend of abstracted groundwater sources identified by Barlow and Leake (2012) show that groundwater storage and surface-water capture together contribute to the total amount of abstracted groundwater. In this study we distinguish a total of three distinct sources of abstracted groundwater (Figures 2.1 and 2.2, Table 2.1)

(1) **Shallow storage** (*SHALGW*): groundwater from storage release in the shallowest part of the system (<100 ft below ground surface). In the numerical model this is calculated as the change in storage volume of all the cells in model layer 1 within the local area, that is, to a depth no greater than 100 ft below land surface.
(2) **Deep storage** (*DEEPGW*): groundwater from storage flowing into layer 1 from deeper

model layers, calculated in the model as the change in volumetric flux from layer 2 to layer 1.

(3) **Surface water** (*SURFW):* change in volumetric flux between surface water features and the aquifer, representing water captured from surface water sources. SURFW is the sum of water coming from streams (streamflow routing cells), lakes and wetlands (drain cells)

The choice for splitting up the contributions for shallow storage (SHALGW) and deep storage (DEEPGW) comes from the fact that we simulate groundwater abstraction from an unconfined shallow aquifer. Ideally and theoretically, these aquifers are not connected to other (deeper) aquifers, but over long abstraction times this cannot be ruled out. In addition, we can test whether ANNs are useful for some or all contributions of pumped groundwater.

The predictor parameters used in this study all represent easily obtainable field and model parameters with some causal relation to the response variables, such as the minimum distance of the well to the nearest stream, hydraulic diffusivity of the unconfined aquifer, average recharge rate etc. (Figure 2.1, Table 2.1).

Table 2.1 Description of predictors and responses used in this study. The predictors in italics and marked by * were dropped after the ANCOVA analysis (section 2.3.3) due to lack of statistical relevance to the response parameters.

Parameter	Description
PREDICTOR (10)	
MIN_DIST	Minimum distance of the seeded well to the nearest active stream
N_EXIST	Number of existing (non-seeded) wells within the local area (-)
LOG_Kh*	Log of the mean horizontal hydraulic conductivity of the local area,
LOG_Kv*	Log of the mean vertical hydraulic conductivity of the local area
STREAM_DENS	Stream density (%). Defined as the number of active stream flow routing (SFR) cells over the total number of cells in the local area
WETL_DENS	Wetlands density (%). Defined as the number of drain (DRN) cells over the total number of cells in the local area
HDIFF	Mean hydraulic diffusivity, defined as the mean hydraulic conductivity over the mean specific yield of all cells within the local area
RECHSUM	Sum of recharge within the local area (ft^3/day)
AQ_THICK	Mean aquifer thickness (ft)
ELEV_DIFF*	Total difference in elevation within the local area (ft)
RESPONSE (3, each at five output times [0.5 yr, 1 yr, 5 yrs, 10 yrs and 25 yrs)	
SHALGW	Source water coming from storage in the pumped layer (layer 1)
SURFW	Source water coming from surface water features (wetlands and/or streamflow
DEEPGW	Source water coming from storage in deeper layers (layer 2 and below)

2.2.2 Numerical groundwater models

For this study we built a total of seven databases built on numerical model data in cooperation with the USGS (Feinstein et al., 2018). The groundwater models used in this study were developed by the USGS and are located around Lake Michigan in the United States (in the states of Michigan, Illinois and Wisconsin, Figure 2.2). The models are based on the MODFLOW-NWT code groundwater models (Niswonger et al., 2011). The models represent the Kalamazoo watershed (KALA) in Michigan, the Upper Fox River watershed (UPFOX) and the Manitowoc-Sheboygan (MANI) watershed in Wisconsin, and were originally studied for research determining groundwater age, but have been adapted for this study to make them suitable for transient groundwater abstraction simulations. All watersheds represent 8-digit hydrologic unit code (HUC8) basins. Model development is explained in greater detail in the report by Feinstein et al. (2018). Here, only the basic model structure and adjustments to the models for this study are described.

Table 2.2 Model structure parameters of MODFLOW-NWT models, specific yield and storage coefficients used based on model hydraulic conductivity

Parameter	**KALA**	**MANI**	**UPFOX**
Number of rows	730	910	1050
Number of columns	1230	530	430
Number of layers	8	15	15
Cell size (ft)	500	500	500
Total surface area model (km^2)	2.09×10^4	1.12×10^4	1.04×10^4
Hydraulic conductivity [ft/day]	**Specific yield [-]**	**Specific storage [ft^{-1}]**	
$K_h < 1$	0.18	2.5×10^{-4}	
$1 \leq K_h < 10$	0.20	1×10^{-4}	
$10 \leq K_h < 100$	0.22	5×10^{-5}	
$K_h \geq 100$	0.25	1×10^{-5}	

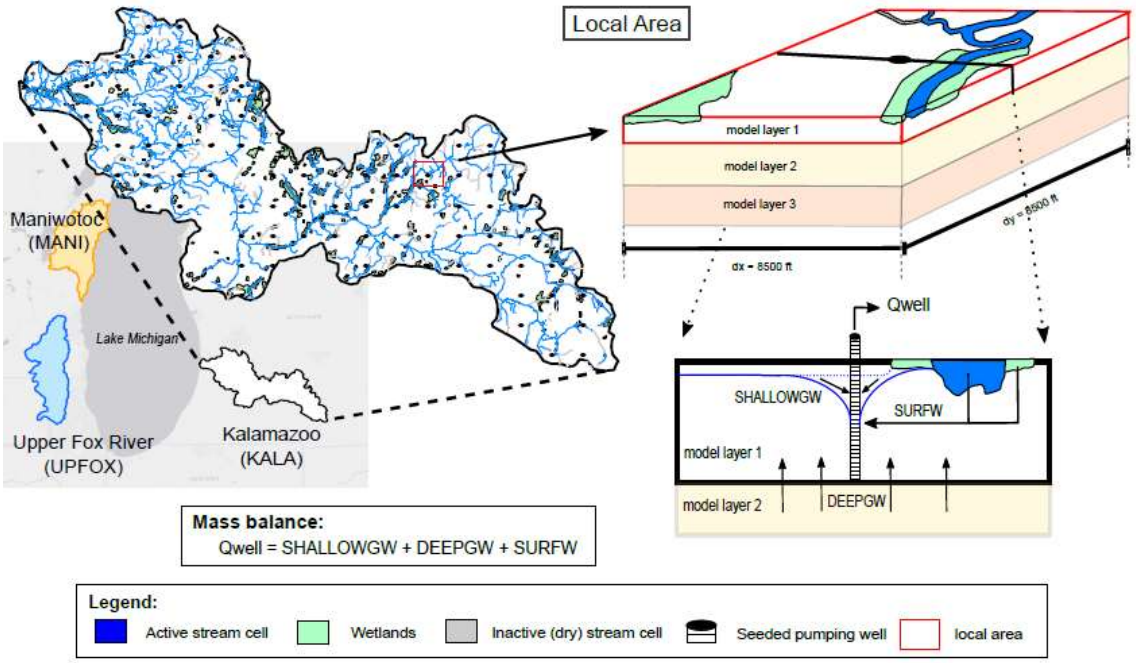

Figure 2.2 Schematic representation of numerical groundwater models (MODFLOW) and local area approach. The figure shows the outlines and locations of the Upper Fox River (UPFOX), Manitowoc (MANI) and the Kalamazoo (KALA) watersheds in the United States. A zoomed-in view of the KALA numerical model with the stream network, wetland areas, seeded abstraction wells and the outline of one local area (red square). In the top right, a box representation of a local area within the numerical model is given with a cross section across the local area under it (bottom right)

To adapt the models from steady state to transient groundwater flow, we initially assigned specific yield (S_y) and storage coefficient (S_s) values based on Domenico & Schwartz (1998) (Table 2.1) for all layers in all models. A sensitivity analysis was performed to determine how the water table would change based on changes in S_y and S_s. As expected in the absence of other stresses, the changes were found to be negligible, and it was therefore concluded that the models were not responsive to these values.

The streams within the basins were represented by Stream-Flow Routing (SFR) package in MODFLOW (Niswonger & Prudic, 2005). Drain cells (DRN) in the model represent lake and wetland areas. The advantage of the SFR package over the use of the more commonly used River Package is that the SFR package allows streams to run dry if there is not enough groundwater discharge in the reach and upgradient to sustain channel flow.

The models were run twice under transient conditions: one simulation where groundwater abstraction by a set of *seeded* (simulated) abstraction wells is simulated and a 'base run' where no groundwater abstraction by seeded wells is simulated. The difference between the two runs is the effect on the groundwater flow system by the implementation of a single groundwater abstraction well. Models were run for 1 stress period under transient conditions with stress period lengths of 1 month, 6 months, 1 year, 5 years, 10 years, and 25

years. The results at these output times are then used to model the overall trend of depletion (storage and surface water) over time (Figure 2.1).

We trained our ANNs based on seven distinct databases. Three databases contain only samples from one single HUC8 basin (KALA, MANI or UPFOX), three databases contain samples from two HUC8 basins combined (KAMA (KALA-MANI), MAUP (MANI-UPFOX) and KAUP (KALA-UPFOX). The final database (KAMAUP) contains data from all the HUC8 basins combined.

2.2.3 Local area definition

To build the databases, we simulated the response of surface water, shallow groundwater, and deep groundwater to hundreds of *seeded abstraction wells* arranged in a grid around each domain (Figure 2.2). To avoid having to run each model hundreds of times, we used the *local area* approach to define a grid of abstraction wells which would minimally interfere with each other in each simulation (Feinstein et al., 2016; Fienen et al., 2016). Each local area is a square subdomain surrounding a seeded abstraction well which is assumed to be independent of all other abstraction wells, allowing us to simulate multiple abstraction wells (with a constant abstraction rate Q_{well} of 25000 ft^3/day) in a single simulation (Figure 2.2). Shifting the grid of the abstraction wells over the domain changes the location of each synthetic abstraction well and local area and therefore represents a new sample to populate our database. One simulation produces between 60 to 100 local areas dependent on the model. By shifting the grid of wells 40 times, between 2400 and 4000 samples are obtained for each domain. Predictor parameters are then calculated for each local area independently based on the model setup.

The methodology of using local areas as independent samples for the ANN relies on the assumption that the areas are sufficiently distant, so abstraction only affects within each local area. To ensure that the local areas were independent of each other, we conducted a series of experiments changing the well spacing and abstraction rate to find the optimal local area and abstraction rate for our simulations (supplementary information). *Well spacing* is defined as the number of model cells between seeded wells. *Local area size* is defined as the number of model cells on one side of a local area (including the cell with the abstraction well). Local area size is always smaller than well spacing since this allows for a buffer zone between local areas.

A mass balance approach (Figure 2.2, Equation 2.1) is taken to determine the most efficient seeded well spacing and local area size. Since two model simulations were run (with and without abstraction), the difference in fluxes between the simulations is the effect of one additional abstraction well (with abstraction of Q_{well}) for the local area.

$$SHALGW + DEEPGW + SURFW = Q_{well} \tag{2.1}$$

Local areas where $SHALGW + DEEPGW + SURFW$ was not within ± 5% of Q_{well} throughout the entire simulation (up until 25 years), were removed from databases. If this mass balance threshold is met, it indicates that sources of water sources outside a seeded well's local area did not significantly contribute to the water abstracted by that seeded well, confirming the assumption that the local areas are independent. The main reason for the mass balance threshold to not be met was a significant contribution of lateral groundwater flow into the local area. Changes in lateral flow in/out of the local area were found to be negligible for most local areas when using a well spacing of 50 cells (25,000 ft) and a local area size of 27 cells (13,500 ft) and therefore this well spacing and local area size was chosen in this study, which is also equal to the local area size used in the study by Feinstein et al. (2016).

2.2.4 Artificial neural network training and accuracy

Once the databases of the different basins have been built, they can be prepared for training of the ANN. Since all input parameters have different units and parameter values can differ by several orders of magnitude, we scaled the input parameters to reduce the learning time for the ANN by computing the z-score for each predictor variable (equation 2.2):

$$x_{norm}^{(i)} = \frac{x^{(i)} - \underline{x}}{\sigma_x} \tag{2.2}$$

Where $x_{norm}^{(i)}$ is the normalized value (z-score) of sample i in the dataset for input feature x, $\underline{x}$ is the mean value for input feature x over all the samples in the dataset and σ_x is the standard deviation of predictor x over all samples in the dataset. For computing the z-score of the combined database (KAMA, MAUP, KAUP and KAMAUP), we combined the data of the individual databases (KALA, MANI and UPFOX) together prior to computing the z-score.

The databases were split into a training set (90% of all samples) on which the ANN is trained and a test set (10% of all samples) which are used to make predictions on the trained ANN (and not used in the training process). The training set is further separated into training data (80%) and validation data (20%). This process has been summarized as a flowchart in Figure 2.3. We performed K-fold cross validation with $K = 10$ to avoid any biased splitting of the database. In K-fold cross validation the splitting of the database into a training set and test set is repeated randomly and differently K times. Therefore, ANNs are also trained K times. The overall skill (section 2.5) of the ANN is determined by the mean skill over all K ANNs. The best databases for ANN training are given with negligible differences in skill between these ANNs.

Finding the optimal ANN structure was done in sequential order. We started with testing suitable cost functions and activation functions in a small 2-hidden layer ANN with 20 nodes per hidden layer. We used all seven remaining predictor parameters for our initial simulations. We found that 6 hidden layers of with 100 nodes each were suitable to model the responses for shallow groundwater (SHALGW) and deep groundwater (DEEPGW). The response variable

surface water (SURFW) was difficult to predict (see Results and Discussion section). The number of hidden layers, the number of nodes per hidden layer, optimizer, cost functions and the activation functions were kept constant for all simulations. The number of epochs (the number of times the entire training date is 'fed' to the ANN) and learning rate were the major hyper parameters used for ANN tuning (depending on the response variable and database used). The Python library Keras (Gulli & Pal, 2017) was used for building the ANN.

2.2.5 Predictive skill of the artificial neural network

We used two different metrics for defining ANN skill. Since we predict a trend rather than a single continuous value (Figure 2.2), we looked at how the ANN is able to predict at specific output times (t_1 through t_5) or as the trend as a whole.

We defined the predictive *skill* of the ANN (also called its accuracy) therefore in two ways. We used the coefficient of determination (R^2) between the predicted ANN value for response i, $y_{pred}^{(i)}$ and the actual value in the database ($y^{(i)}$), which is output from the numerical groundwater models. We computed the coefficient of determination at each individual output time (and therefore output node). Therefore, we computed five values of $R_{t_i}^2$, where t_i is the output time. These five $R_{t_i}^2$ values depict how good the ANN is performing at parts of the trendline.

The overall ANN skill is given by the mean of all $R_{t_i}^2$, where n is the number of output nodes and $K = 10$ (for using 10-fold cross validation).

$$R_{NN}^2 = \sum_{i=1}^{K} \frac{\sum_{i=1}^{n} R_{t_i}^2}{n} \tag{2.3}$$

In addition to the coefficient of determination we calculated the mean squared error (MSE) and mean absolute error (MAE) for the optimal predictor parameter sets to check for bias. These were found to be comparable to the results using the coefficient of determination metric.

2.2.6 Predictor optimization strategy

Even though the proposed predictor parameters are causally connected to the response variables (from a hydrological standpoint), this does not mean that they are suitable for use in an ANN. An ANN is a purely mathematical and statistical tool, which by itself does not consider any physics-based relations. Including predictors which have a physics-based relation to the response variable but lack any statistical relation might cause the ANN to underperform. The same effect could occur when predictors are used which give either duplicate or contradicting pieces of information.
In this study we attempted this parameter optimization approach in a hybrid manner: we used hydrological expertise combined with a statistical and 'brute force' approach to

systematically optimize our parameter selection. The proposed procedure includes:

Step 1: Make selection of predictor parameters with a causal relation to the expected decline of surface water and groundwater storage due to pumping (including parameters used in analytical solutions and numerical models).

Step 2: Perform a statistical Analysis of Covariance (ANCOVA) for determining statistical one-on-one relation between the selected parameters and the relevant output parameters. The ANCOVA test combines properties of ANOVA and multiple linear regression to ensure that the predictor parameters used in this study 1) are statistically relevant to the response variables and 2) are unique and independent of other predictors. In this way we can not only estimate the relevance of each individual predictor to each response variable, but also the relevance of each individual predictor x_i when taking in combination with another predictor (for example $x_i : x_j$). The methodology used in this study is similar to the one used by Li et al. (2014) and the tests were performed in R (Ihaka and Gentleman 1996).

A P-score smaller than 0.05 implies the predictor parameter is determined to be statistically relevant the response variable. We performed the ANCOVA test for all response (SHALGW, DEEPGW and SURFW) at three output times (t_1 = 0.5 yr, t_3 = 5 yrs and t_5 = 25 yrs). The ANCOVA test process is summarized in Figure 2.3.

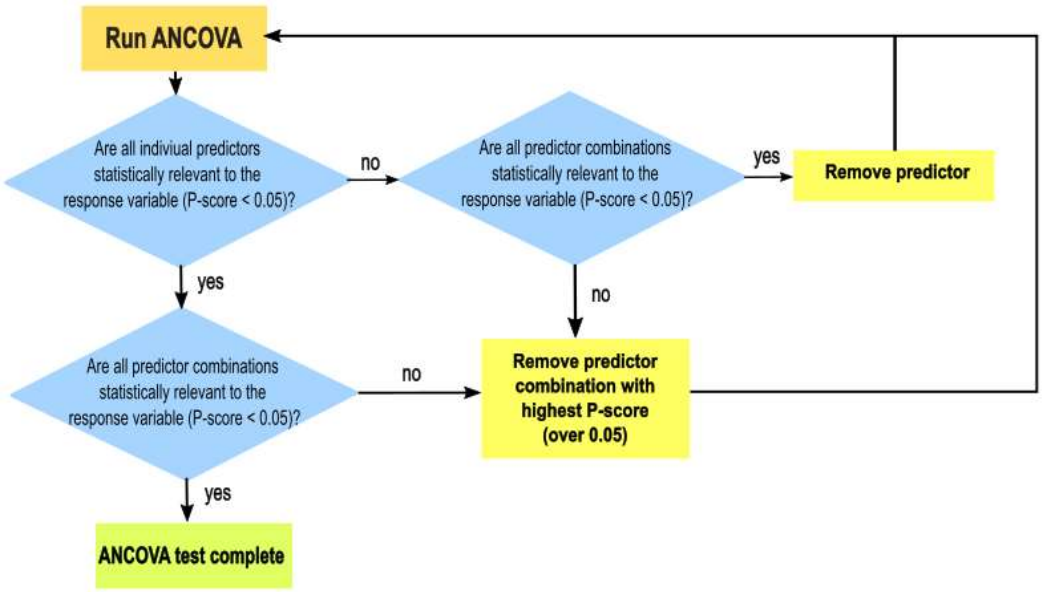

Figure 2.3 Flow chart of ANCOVA process for determining statistical relevance of the predictors to the response variables. The ANCOVA process was repeated for every response variable (SHALGW, DEEPGW and SURFW) and for all databases.

With this process we were able to eliminate three predictor parameters (LOG_Kh, LOG_Kv and ELEV_DIFF) from our predictor variables, as the parameters were not found to be statistically relevant to any response variable during any output time in any database.

RECHSUM was found to be statistically irrelevant to the response variable SURFW, but relevant to the other response variables (SHALGW and DEEPGW). Therefore, this predictor was retained. The matter of predictor redundancy is tackled in the predictor optimization process (section 2.6).

Step 3: The final step in the parameter selection process is to minimize the number of predictors to test the sensitivity to number of predictors and to filter out any redundant predictors. We started out with a predictor parameter set with all seven predictor parameters which is assumed to be our base model (Table 3). Sequentially, we trained ANNs with only six predictor parameters, where in each simulation a different predictor parameter was omitted from the base model. The results (R^2_{NN}) of these ANNs with six predictor parameters are then compared to our base model. If the R^2_{NN} value of the ANN with six predictors does not change compared to our base model (or even increases), then the predictor omitted from this ANN since it did not contribute optimally to the predictive skill. The predictor corresponding to the ANN with the highest R^2_{NN} is then excluded. The remaining six parameters are taken over and this process is repeated for ANNs with 5, 4, 3, and 2 predictor parameters. For the sake of simplicity, ANN structure (number of hidden layers and nodes) was kept constant throughout the simulation, but the number of epochs and the learning rate were varied to obtain optimal learning conditions for the ANNs.

Table 2.3 Predictors used in the base model for artificial neural network training

	Predictors in base model
1	WETL_DENS
2	N_EXIST
3	STREAM_DENS
4	MIN_DIST
5	RECHSUM
6	HDIFF
7	AQ_THICK

2.2.7 Cross-basin comparison

Ultimately, ANNs could potentially be used to produce proxy data for remote regions where field data is scarce and therefore numerical modeling is not a feasible option. We have simulated this by using a 'round-robin' format of using the combined databases (KAMA, MAUP, KAUP and KAMAUP) as our trained databases and using testing data from the HUC8 basins not used in training (UPFOX, KALA, and MANI respectively).

In addition, we looked at the other possibility if a combined database can also be used when only testing data from one part of the model domain is used. For this we used all the combined databases, including KAMAUP and used testing data from all possible sub domains: either KALA, MANI or UPFOX individually and KAMA, MAUP and KAUP for the KAMAUP database.

We used our base model as our predictor parameter set. Since the predictor range differs from database to database, we normalized the testing data over the mean and standard deviation of the database used for training. Therefore, we used an adapted version of equation 2.4:

$$x_{norm}^{(i)} = \frac{x_{test}^{(i)} - \underline{x}_{train}}{\sigma_{x,train}} \tag{2.4}$$

Where $\underline{x}_{train}$ and $\sigma_{x,train}$ are the mean and standard deviation of the predictor x in the training database and $x_{norm}^{(i)}$ is the normalized value (z-score) of sample i in the dataset for input feature x from the testing data. This reduces the penalty for having predictor values in the test set which are outside of the domain of numbers in the training set.

2.3 Results and discussion

2.3.1 Model results

Figure 2.4 shows the model results for the individual contributions of the three response parameters per basin and output time. The results show the mean cumulative contribution of the response parameter over all local areas in the dataset as a percentage of total abstraction rate ($\frac{Q_{source}}{Q_{well}}$).

Storage release remains the dominant source of abstracted groundwater throughout the entire simulation time, while surface water is found to be a minor contributor with up to 11.2% of total source water storage release from shallow parts of the aquifer (SHALGW) decreases from ~85% at shorter abstraction times to ~50% at longer abstraction times. The KALA basin shows the lowest contribution of SURFW (up to 4.5% after 25 years), while the UPFOX basin shows the highest contribution of SURFW (up to 11% after 25 years).

We believe the low numbers of surface water contribution (SURFW) can be explained by the relatively low streamflow numbers in the numerical models. This could be different when using a different balance of abstraction rate and local area size. However, we kept the current abstraction rate and local area size since this allows us to compare the optimal predictor selection to the work by Fienen et al. (2016), since they used comparable values for abstraction rate and local area size.

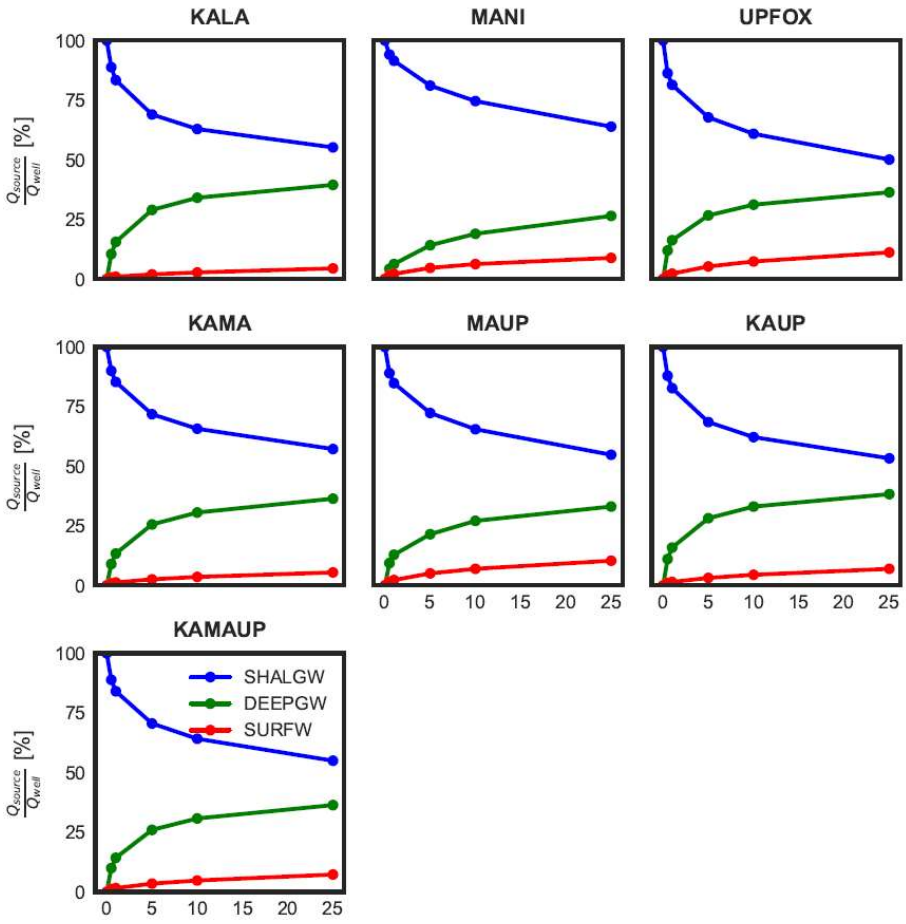

Figure 2.4 Break down of the mean cumulative contribution of all three response parameters (SHALGW, DEEPGW and SURFW) over all local areas in the databases. Results are given for all databases and output times.

2.3.2 Artificial neural network agreement with numerical models

The overall ANN prediction performance (R^2_{NN}) for each database and response variable is given in table 2.4. Maximum test R^2_{NN} was found to be 0.82 for SHALGW and 0.84 for DEEPGW with values up to 0.95 for correlation at individual time steps ($R^2_{t_i}$, Figure 2.5). For the response variable SURFW, the ANN skill was found to be too low to be useful for prediction purposes, with a maximum test R^2_{NN} of 0.52. This is primarily caused by the low numbers for streamflow, especially at low abstraction times. Figure 2.5 shows the results of ANN training on the base model for all databases (KALA, MANI, UPFOX, KAMA, MAUP, KAUP and KAMAUP) for every output time ($R^2_{t_i}$) and there it can be seen that the ANN skill at each response time improves for SURFW as the overall contribution of SURFW to $\frac{Q_{source}}{Q_{well}}$ increases (up to $R^2_{t_5}$ = 0.70 for the MANI database).

Optimal ANN skill for both SHALGW and DEEPGW is obtained at the earliest and latest output times (up to 0.96). Therefore, the ANN performs equally good for both response variables.

Table 2.4 Artificial neural network performance for each response variable and each database using the base model (table 2.3)

Database	SHALGW			DEEPGW			SURFW		
$N_{pred} = 7$	Train R2	Val R2	**Test R2**	Train R2	Val R2	**Test R2**	Train R2	Val R2	**Test R2**
KALA	0.85	0.81	0.82	0.90	0.84	0.84	0.52	0.37	0.25
MANI	0.81	0.79	0.77	0.93	0.85	0.83	0.74	0.67	0.38
UPFOX	0.85	0.80	0.79	0.90	0.81	0.81	0.56	0.5	0.52
KAMA	0.82	0.81	0.82	0.88	0.84	0.83	0.62	0.51	0.51
MAUP	0.85	0.79	0.78	0.90	0.82	0.82	0.75	0.59	0.51
KAUP	0.8	0.79	0.79	0.88	0.82	0.83	0.69	0.48	0.45
KAMAUP	0.84	0.80	0.80	0.87	0.82	0.82	0.64	0.39	0.26

2.3.3 Predictor parameter optimization

Figure 2.6 shows the results of the predictor optimization process explained in section 2.6. We trained ANNs with lower numbers of predictors (up to a minimum number of 2 predictors) to predict the responses DEEPGW and SHALGW. We did not repeat this process for SURFW in this study, since we already found insufficient correlation in the base model, and it was hypothesized that the highest ANN skill would be obtained at a higher number of predictors. The results show that the ANN is generally good at predicting the contribution simulated by the physics-based model of shallow groundwater (SHALGW) and deep groundwater (DEEPGW) for predictor parameter sets of length 5 through 7, where the maximum ANN skill of 0.84 is obtained. Optimal ANN skill decreases to a 0.6 – 0.7 range when only 2 predictors are used. The overall ANN skill is lower compared to other related studies where non-deep learning models were used (Ahmad & Simonovic, 2005; Antonopoulos et al., 2016; Daliakopoulos et al., 2005; Zealand et al., 1999), but comparable to other recent studies (Afzaal et al., 2020). This is caused partly by the more complex nature of our methodology which is different from other studies. Other neural network studies used a time-lagged effect for groundwater depletion or groundwater level fluctuations. This can be combined with a transient dataset in climate parameters to calculate the groundwater head at certain time steps.

We found differences and similarities in the predictor optimizing process when using the different databases for training. The annotations in Figure 2.6 depict the predictor that was dropped after each subsequent reduction in the number of predictors used in training the ANN (for example in the KALA model, stream density (STREAM_DENS) was found to be the least important predictor and was dropped when reducing the number of predictors to 5, etc.). The predictors that were dropped are minimum distance (MIN_DIST), stream density (STREAM_DENS), wetland density (WETL_DENS), number of existing wells (N_EXIST) and/or recharge sum (RECHSUM), albeit in a different order depending on the database used. On the

other hand, we found that hydraulic diffusivity (HDIFF) and mean aquifer thickness (AQ_THICK) were found to be the most important to obtain high ANN skill. We didn't find notable differences in predictors for predicting SHALGW or DEEPGW.

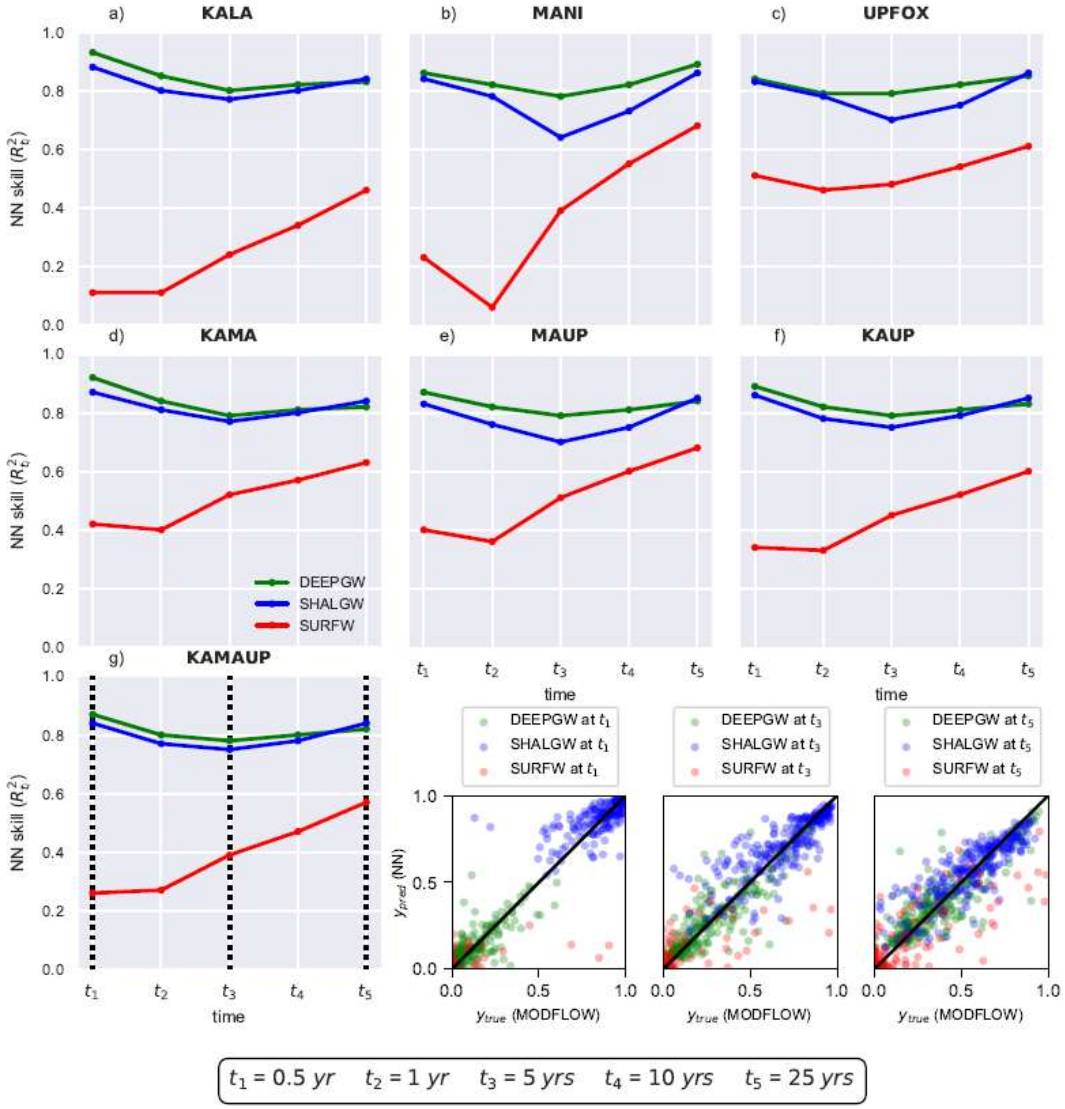

Figure 2.5 Artificial neural network skill for artificial neural networks trained on the seven distinct databases. The artificial neural networks are trained on the base model (with seven predictors). Graphs A through G show the artificial neural network skills for artificial neural networks trained on the A) KALA, B) MANI, C) UPFOX, D) KAMA, E) MAUP, F) KAUP and G) KAMAUP database. The three scatter plots show the correlation of the predicted (artificial neural network) values (y_{pred}) and the value of th numerical (MODFLOW) model (y_{true}) on which the coefficient of determination is estimated. The three scatter plots show the correlation of the three response variables (SHALGW (blue), DEEPGW (green) and SURFW (red) at three output times (t_1= 0.5 yrs, t_3= 5 yrs and t_5= 25 yrs), corresponding to the vertical dotted lines in graph G (KAMAUP database)

Predictor parameter sets where HDIFF was omitted produced the least performing ANNs regardless of database used. The differences in variance (2 standard deviations over 10-fold

cross validation, 2σ) between the KALA/KAMAUP and MANI/UPFOX results contributed to the differences in size of the databases. MANI and UPFOX are the smallest database and therefore it is expected to have the largest variance over 10-fold cross validation.

Fienen et al. (2016) concluded that recharge and transmissivity were not suitable predictors and found better agreement when those predictors were not included in the predictor set. Under steady state conditions, this makes sense from a hydrological point of view since recharge is assumed constant throughout time. In this study we found that the inclusion of transient effects and other sources of water (primarily storage release) changes the dynamics and increases the complexity of the problem and that recharge and transmissivity (in the forms of predictors RECHSUM and HDIFF) are important for ANNs. Under transient conditions predictors themselves are transient as well, with changes between 5-25%. For example, we calculated surface water density as the fraction of active streamflow routing cells at the start of the simulation. However, since abstraction leads to decreases in streamflow, surface water density is expected to decrease with abstraction time. Other predictors like MIN_DIST and RECHSUM are also variable with time.

It was beyond the scope of this study to test the possibility of including additional predictor parameters (beyond the predictors in our base model, Table 2.3). We cannot exclude that using different predictor parameters could improve ANN skill. In addition, the inclusion of transient predictor parameters could be used in future studies to improve the ANN model. In this current study, predictors like MIN_DIST and STREAM_DENS were assumed constant through time. This is not the case however in watersheds with many ephemeral streams.

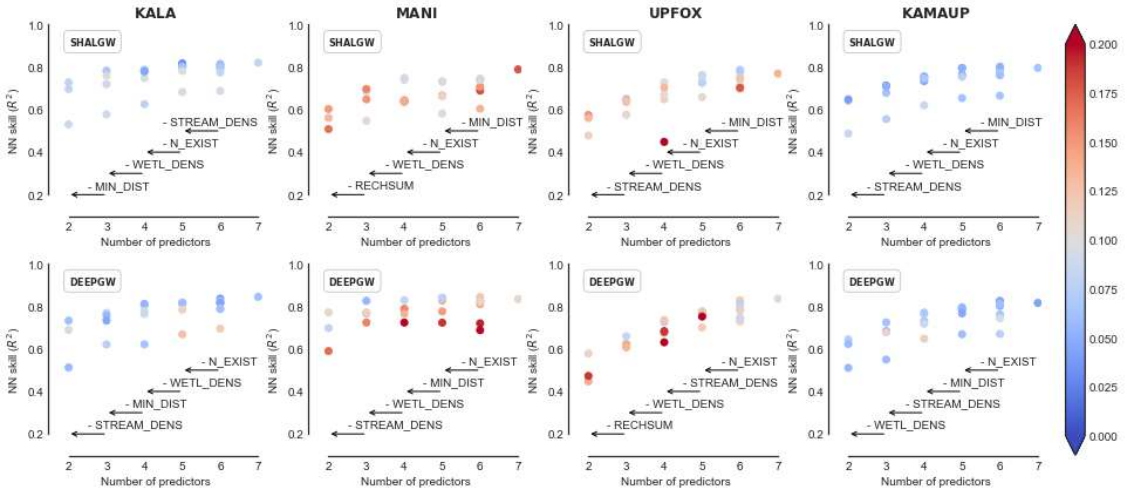

Figure 2.6 Optimal artificial neural network skill (R^2_{NN}) for artificial neural networks trained on 2 - 7 predictor parameters. Artificial neural network skill at $N_{pred} = 7$represents the skill of the base model (Table 2.4 and Figure 2.5). Every independent data point represents an artificial neural network where one predictor is omitted. The annotations depict the predictors that were dropped at each subsequent step, since it was the least contributing or redundant predictor parameter (as is explained in section (2.2.6). The colors represent the range in R^2_{NN}over 2 standard deviations over 10-fold cross validation.

2.3.4 Suitability of numerical models for artificial neural network training

In this study we found that the ANN skill depends on the suitability of a numerical model used for building the database. The numerical models used in this study were suitable when predicting the contribution from storage and therefore are useful when predicting storage depletion. Since the numerical models produced very low values for streamflow, this left most local areas with close to 0% SURFW contribution. The suitability of the numerical models becomes clear when the optimal ANN skill is plotted against the skewness of the response variable at a particular output time (equation 2.5):

$$skewness = \frac{n}{(n-1)(n-2)} \sum \left(\frac{y_i - \underline{y}}{\sigma} \right)^3 \tag{2.5}$$

Here n is the number of samples in the database, y_i is the *i'th* sample in the dataset, $\underline{y}$ is the mean value among all samples and σ is the standard deviation. When we plot the skewness of the response variables in the databases at every output time and plot them against $R^2_{t_i}$ (from Figure 2.5), we obtain the results presented in Figure 2.7. With increasing abstraction time, the relative contribution of surface water (SURFW) to the abstracted groundwater is expected to increase. During low abstraction times (<10 yrs), it is found that for most local areas the contribution of SURFW to abstracted groundwater is negligible. When the database consists of a majority of zero values compared to only a few non-zero values, it becomes difficult for an ANN to train on. We found that the database where the response variable has a skewness in the range of [~ -2.5, ~ +2.5] can be used for training. When the data becomes more skewed than this threshold value (as is the case for SURFW) then the "quality" of the database becomes too poor for an ANN (with our currently used structure) to be trained. ANN skill could still be improved by changing the overall structure and tuning hyper parameters, but that was not within the scope of this study.

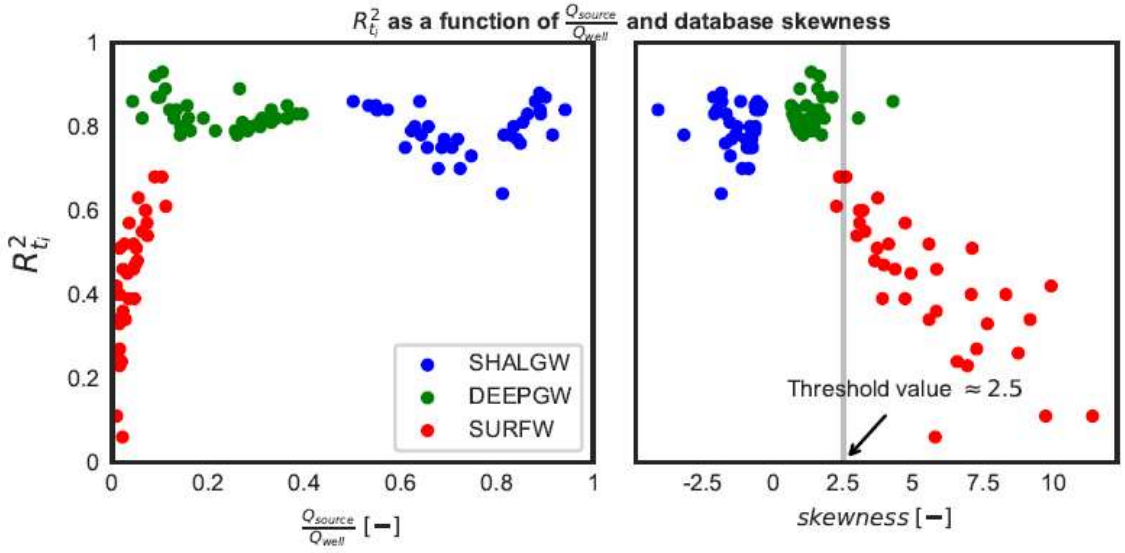

Figure 2.7 Correlations of relative contribution of response variable to the total abstraction rate ($\frac{Q_{source}}{Q_{well}}$) and database skewness against neural network still (R^2_t). The three colors represent the three response variables.

2.3.5 Using ANNs over conventional methods

The purpose of using an ANN over a conventional numerical groundwater model is that the ANN uses input parameters which are site specific and easy to obtain, and computations can be performed rapidly compared to numerical models, allowing for quick testing of many different scenarios. While ANNs have been used to estimate surface water depletion before (Fienen et al., 2016), this is among the first studies where we incorporated an optimization process for estimating redundant predictor parameters.

We argue that even though the results of this study are promising, we did not find that the trained ANNs outperformed the corresponding numerical model simulations. We contribute this to the use of 'local areas' as individual samples which do not seem to work under transient conditions (this study) compared to steady state conditions (previous studies).

This study shows that an ANN with five to seven predictor parameters predicts storage depletion with a correlation of ~0.84 compared to a numerical groundwater model. Whether this accuracy is satisfactory as a 'first-order approach' or as a substitute for numerical models is still up for debate for hydrologists or water managers to decide. Since these predictors parameters are easily obtainable from field measurements or satellite imagery, making predictions using the ANN can be performed faster and easier than the complex construction of a numerical model. The prediction holds for when using an ANN trained on a database from a single HUC8 basin (for example KALA), but also for database containing data from multiple basins (KAMAUP). The low variance (2σ = ± 0.03) in correlation shows that the prediction is robust when the database is large enough.

Others might disagree with the need of a trained ANN on modeled data since the numerical model was built in advance of training the network. An ANN would not be an asset as a surrogate model for the numerical model. However, there are scenarios where an ANN trained on modeled data could be useful. Numerical MODFLOW models are only built and calibrated for specific geographic locations and specific time periods. ANNs can be trained on numerical model data trained on time series data of groundwater levels, fluxes, or surface water levels within the domain of the calibration period and used to predict future groundwater levels etc. In this way ANNs can serve in combination with numerical models. We attempt to test this in future work in addition to improving our methodology.

2.3.6 Using ANN as a proxy for data from other regions

The results of our cross-basin comparison are displayed in Figure 2.8. We found that ANNs trained with any combined database were 1) able to accurately predict test data from within its combined domain and, 2) able to accurately test data coming exclusively from a sub-domain of total domain of the training data; e.g. an ANN trained on the KAMA database is able to correctly predict KAMA test data, which is the combined domain of both the KALA and MANI HUC8 basin, but also test data from only one of the two basins. Performance drops when data from outside the domain of the training set is used. This is shown in Figure 2.8a-c. The scatterplots represent the ANN prediction against MODFLOW data for an ANN trained on the KAMA database, while test data is used from the a) KALA, b) KAMA and c) KAMAUP database. The individual data points are marked with the origin (HUC8 basin) of that data point. The three plots show three possible scenarios:

1. Subplot b) represents the scenario where test data is used from the entire domain of the training data.
2. Subplot a) shows the scenario when only test data from a sub domain or selective area of the total model domain is used (in this case the KALA HUC8 basin). The ANN skill for this scenario is comparable to scenario 1 (subplot b).
3. Subplot 3) shows the scenario where test data is used from outside the original training data domain. Test data in this case comes from the KAMAUP database, which is a combination of data from all HUC8 basins. Data from the UPFOX basin is not used in training.

The results for scenario 3 show that the lower ANN skill is (as expected) mostly due to the inclusion of UPFOX data. Most of the outliers in the prediction represent data points from the UPFOX basin on which the ANN has not been trained on. This also explains why the KAMAUP database shows the overall best performance when using test data from sub-regions of the overall training domain. Since the KAMAUP database is trained on the full range of predictor parameter values over all sub-domains, it can predict for all sub-domains. When the predictor values of a test sample fall far outside the range of the training data, then the ANN falls short in predictive power. Based on our current methodology we conclude that although the first impressions are promising, the ANN lacks the performance power as a predictive model for cross-basin comparison.

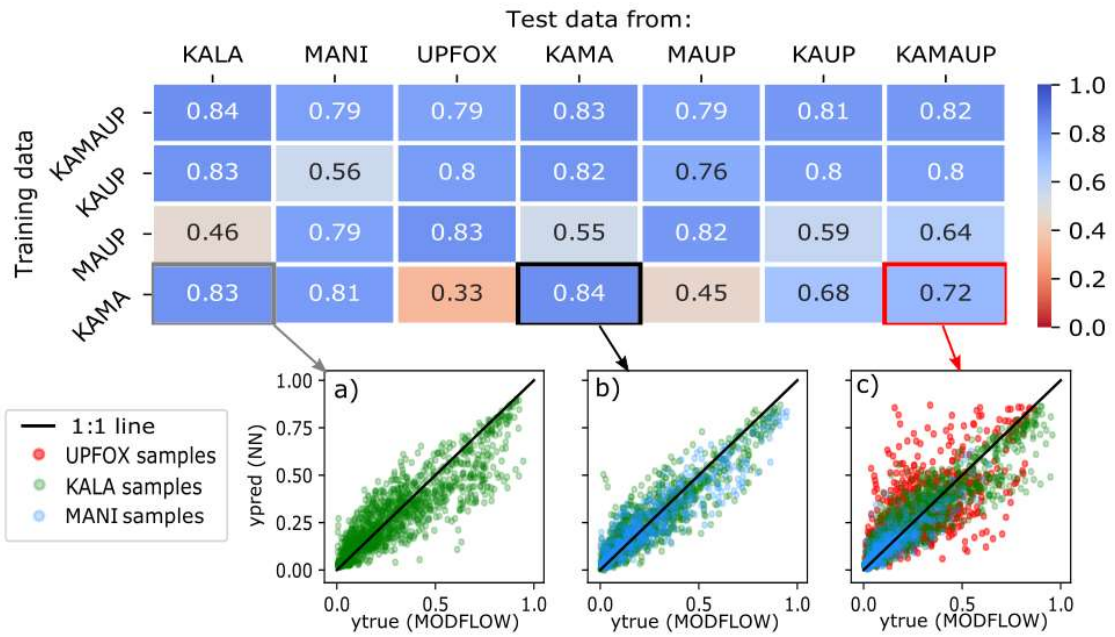

Figure 2.8 Heat map showing the ANN predictor skill $(R^2_{ANN})$for cross-basin comparison averaged over 10-fold cross validation. Training datasets are given on the vertical axis and testing data is shown on the horizontal axis. Subplots a), b) and c) represent scatter plots for an ANN trained on the KAMA database and test data used from the a) KALA, b) KAMA and c) KAMAUP database. The samples are marked by database origin (KALA: green, MANI: blue and UPFOX: red).

2.4 Conclusions

In this study we tested the ability of ANNs to predict sources of water to wells caused by groundwater abstraction under transient conditions. The results show that the ANN can predict the different possible sources of abstracted groundwater with a maximum agreement (R^2) to the underlying physics-based model of 0.84 (2σ = ± 0.03), depending on the source (shallow groundwater, deep groundwater, or surface water). ANNs were found to be underperforming when predicting the contribution of surface water (up to R^2 = 0.6). This is contributed to the low values of streamflow in the numerical models.

ANNs perform better when between 5 and 7 predictors are used and when the size of the training database increases. We found that hydraulic diffusivity, mean aquifer thickness and recharge are the most important predictors to include for predicting storage depletion under transient conditions. This contrasts with previous studies where ANNs were used to predict sources of water to wells under steady state conditions.

We also explored the possibility of ANNs in generating proxy data. The results show that ANNs are usable for prediction purposes when test data originates from within the model domain used for training. This includes the hypothetical case where test data from only a sub domain of the total training domain is used. However, the ANN fails to perform when we attempt cross-basin prediction. This is attributed to a combination of the numerical models used in this study as well as the methodology, which should be improved upon.

Overall, this research has shown that ANNs can be beneficial in addressing hydrological and environmental problems. simulations. ANNs can be used to predict storage depletion over time using only a few easily obtainable parameters with comparable results to a numerical model, but they have not outperformed numerical models. Future and ongoing research will test whether improvements to this methodology can make the models more robust.

Chapter 3: Disentangling process controls on global groundwater table depth patterns using random forests

3.1 Introduction

Groundwater is the dominant source of freshwater for human consumption, irrigation and for industrial purposes (Wada et al., 2010; Wada & Bierkens, 2014). However, the availability of freshwater in the future is coming under increasing stress because of global population growth, climate change, and depleting water resources. Global modeling of groundwater, recharge flow and depletion is useful for estimating global groundwater availability (Gleeson, et al., 2012; Gleeson et al., 2016) and for good estimations of the planetary boundary of fresh water (Gerten et al., 2013; Gleeson et al., 2020; Rockström et al., 2009). A recent perspective paper by Gleeson et al. (2021) argued the four salient reasons for regional and global scale groundwater modeling are: 1) To understand and quantify past, present and future groundwater-climate connections, 2) to understand and quantify interactions between groundwater and water in other parts of the hydrologic cycle, 3) to help inform (transboundary) water policy decisions and 4) to create visualizations and interactive opportunities that inform citizens and water consumers.

A global model for determining water table depth would help improve our understanding within the framework of these four reasons. For that matter, regional and global models are useful to determine potentially vulnerable regions where groundwater availability might be insufficient, either currently or in the future. However, a limitation of these numerical models is that they often lack sufficient field data in remote regions and are therefore often less suitable making predictions outside their model domain (Reichstein et al., 2019; Wagener et al., 2021). The added difficulty lies in that even though the physics of water flow through a porous medium (soil) is generally well understood, how these physics translate to larger regional or global domains is generally not well understood and the data support is poor (Reinecke et al., 2020). While on local scales, aquifer geometries can be featured quite well in numerical models, questions arise on how to properly present these features for individual model cells on a 1° scale or larger in terms of aquifer heterogeneity and groundwater-surface water interactions among other parameters. We see this in the discrepancies in simulated hydraulic head, water table depth or other fluxes (for example groundwater recharge) physics-based groundwater models (Döll & Fiedler, 2008; Fan et al., 2013; Mohan et al., 2018; Reinecke et al., 2019; Reinecke et al., 2021).

Contrary to physics-based groundwater models (which we define here as numerical models simulating groundwater head or flow on known physical laws of groundwater), machine learning models have emerged as major tools for hydrologic investigation, insight, and prediction. Machine learning is a broadly used term for several mathematical methods (incl.

support Vector Machines, neural networks, gradient boosted decision Trees etc. (Alpaydin, 2014; Bonaccorso, 2017; El Naqa & Murphy, 2015; Hsieh, 2009; Mahesh, 2020)). These methods are trained on known data(sets) and rely more on statistical relevance rather than the "traditional" laws of physics, although recent advances include a hybrid modelling approach where machine learning algorithms are enhanced or enforced by physics-guided loss functions or model setup (Bhasme et al., 2021b; H. V. Gupta, 2019; Kraft et al., 2020). Applications of machine learning within the field of hydrology include streamflow prediction using neural networks (Adnan et al., 2019; Parisouj et al., 2020; Petty & Dhingra, 2018; Rasouli et al., 2012; Shamshirband et al., 2020), transient groundwater table prediction, and groundwater depletion due to pumping (Feinstein et al., 2016; Fienen et al., 2016). One important detail about these studies is that most studies use transient data for their target variable. This is useful since this allows the machine models to predict the target variable considering possible time-delay effects (such as groundwater flow/streamflow). Also, most of these studies focus on making predictions for specific locations (wells or local areas) and don't attempt to predict larger areas or attempt extrapolation to other areas (More, 2018; Zhao et al., 2020). The aforementioned extrapolation problem also persists for machine learning models, since even though the models can make predictions outside the range of their training data, the robustness of their predictions might be unclear (Reichstein et al., 2019).

Building numerical models and machine learning models for the purpose of estimating a variable have similarities, but also major fundamental differences. For physics-based models to predict these target variables to unknown regions they either 1) collect new data from these regions or 2) model these variables using extrapolation of known data (Figure 3.1a). Since the physics of these processes is generally known hydrogeologists are able to conceptualize the relative flow patterns, input variables and boundary conditions into perceptual models (defined as "the evolving understanding of real-world systems based on the interpretation of all available information, influenced by each hydrologist's unique experience and training" (K. J. Beven & Chappell, 2021; Wagener et al., 2021)) which form the base for their numerical model representation of reality. After calibration of their numerical model, scientists are then able to provide their model results to the rest of the scientific community as a basis for future models.

Contrary to numerical model studies, this feedback loop is not so prevalent in comparable studies with a machine learning approach. While physical properties and processes are the core of numerical models; data and mathematical/statistical relevance are the core of machine learning models. This fundamental distinction between these two model types might seem trivial but allows for out-of-the-box ideas to flourish without being constrained to predetermined laws of physics. This perspective allows hydrologists to obtain a perceptual model based on the data, rather than on the assumed physics (Figure 3.1b).

Wagener et al. (2021) proposed an alternative approach to how hydrologists could collect and share information and knowledge on how large scale (groundwater) models operate. Additionally, they noted that perceptual models are underused in current practice

and propose it as a stronger focus within the community. Within the lens of gaining insight on hydrological processes through perceptual models we see potential when we combine this approach with the rapidly growing field of machine learning (Alpaydin, 2014; Bonaccorso, 2017; El Naqa & Murphy, 2015). Beven and Chappell (2021) state that "perceptual models might be useful to put more focus on the understanding of *observable* hydrological processes as a way of improving predictability". In addition, it has been argued that (purely) data-based models (and/or models derived from machine learning) might provide better model representations of reality than models derived from known hydrological processes (K. Beven, 2020; K. J. Beven & Chappell, 2021; Kratzert et al., 2019).

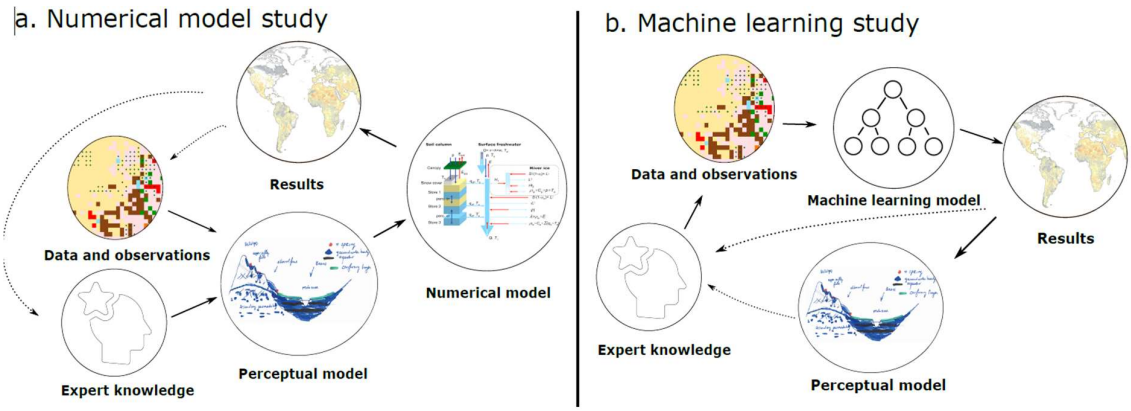

Figure 3.1 Schematic representations of hydrogeological studies involving a) developing numerical models and b) developing machine learning models. Solid arrows indicate steps in the modeling process, while dashed arrows depict feedback loops on how the results are beneficial for gaining scientific knowledge. Icons are obtained from the Noun Project (*Noun Project: Free Icons & Stock Photos for Everything*, n.d.) and Wagener et al. (Wagener et al., 2021)

In contrast to the abundance of research on predicting transient parameters in hydrology, there is not that much research done on predicting non-time dependent (or steady state) parameters. Some recent studies in the field of hydrology narrow down to classification problems, such as a risk assessment of groundwater contamination (Sajedi-Hosseini et al., 2018) or mapping of global groundwater potential (Lee et al., 2020; Prasad et al., 2020). In terms of predicting steady-state non-categorical (continuous) hydrological variables, we think there is underexplored space within the scientific literature as has also been sent in more recent publications (Kratzert et al., 2019; Nearing et al., 2021; Reichstein et al., 2019).

In this study we attempt to fill these knowledge gaps by looking for the potential of generating a global steady-state groundwater table map with a random forest model. We are the first to use the global groundwater table depth dataset by Fan et al. (2013) for this purpose. This dataset consists of over 1.5 million measurements of groundwater table depth obtained from measuring programs from local and federal governments, supplemented with literature data. As input (or *predictor* (Fienen et al., 2016)) parameters for our random forest models we used global (climate) model data sources (Table 3.1). We contribute to the aforementioned

knowledge gaps in three novel and meaningful ways. First, since we predict a steady-state (continuous) parameter, we contribute to the underrepresented array of steady-state regression problems within hydrology. Second, since the Fan et al database has a high global variance in measurement density, it makes it a suitable choice to explore the extrapolation problem within machine learning. Third, the groundwater table depth map generated by machine learning can be compared to groundwater table depth maps generated by physics-based models (De Graaf et al., 2015; Fan et al., 2013; Reinecke et al., 2019) to gain valuable insight in how these models compare.

Using the regional and global groundwater table results, we analyzed, interpreted and contextualized the results to: a) compare the groundwater table depth maps generated by random forests to equivalent maps generated by three physics-based global groundwater models: i) the model by Fan et al. (2013); ii) de Graaf (De Graaf et al., 2015) and iii) G^3M (Reinecke et al., 2019)) which explores whether the physics-based laws of groundwater are also captured by the random forests model; b) Identify differences in perceptual models based on physics-based models and random forests models and c) evaluated the extrapolation potential of our random forest model.

3.2 Methods

3.2.1 Predictors and target variables

There is only a limited known number of studies predicting groundwater table depth using random forests or other machine learning tools (Govindaraju & Rao, 2013) and most often the studies use transient data where water table depth is predicted based on predictor data (Runoff, Precipitation etc.) both for the target time t, but also the predictor input from previous time steps $t-1, t-2, \ldots$etc. This represents the lag time effect of water originating from precipitation/runoff to reach the water table. Under steady-state conditions this lagged time effect is not present and we cannot include this effect. Therefore, we are limited to predictors with a causal relation to the long-term groundwater table depth. We based our selection on expert knowledge and parameters which have been used in previous studies (Bowden et al., 2005; Bowes et al., 2019; Daliakopoulos et al., 2005; Reinecke et al., 2019; Ren et al., 2014; Roshni et al., 2020; Zealand et al., 1999). The predictor data sources are given in Table 3.1. In our selection of data sources, we considered that the data must be open source for our work to be easily checked and to allow our research to be transparent (Bowes et al., 2019; Daliakopoulos et al., 2005; Malik & Bhagwat, 2021).

The final criteria for our selection of predictor parameters is that the parameters show no strong correlation to each other individually. We calculated all pairwise correlations of all predictors of continuous data for all samples in the training data (Table 3.1). The results of this analysis can be found in the supplementary information (Appendix B, Figure B-2). Based on the results we determine that there is no or only small correlation potential between the

predictors in the training data and that we did not choose any redundant parameters. We also considered whether certain predictors should be combined (for example use the aridity index (*P/PET*) instead of precipitation and evapotranspiration separately). We decided to keep these predictors separate for the following reason. It has long been thought that for most catchments' climate aridity is the dominant control of partitioning precipitation into streamflow and evaporation. This has been outlined in the Budyko curve (Budyko, 1951). Gnann et al. (2019) expanded this framework to whether there is also a Budyko curve for climate controls on baseflow. They concluded that in areas with high precipitation, the aridity index can only partly explain the variability in baseflow index (defined as the fraction between baseflow and streamflow). Given the high spatial variance in precipitation and evapotranspiration, we decided to keep these parameters separate.

Table 3.1 Overview of predictor data and target data, spatial resolution of source data and sources. Discrete parameters with (*) can both be considered as both discrete and continuous data, but for this study permeability and porosity have been defined as discrete parameters given the low number of unique values in the datasets (<10).

Predictor	Symbol	unit	Data type	Spatial resolution source data	Model	Source
Surface elevation	***SurfElev***	m	Continuous	1 km	GMTED2010 dataset	(*EarthExplorer*, n.d.), (Danielson & Gesch, n.d.)
Slope	***Slope***	degree	Continuous	1 km	GMTED2010 dataset	(*EarthExplorer*, n.d.), (Danielson & Gesch, n.d.)
Log(Permeability)	***Perm***	m2	Discrete*	1 km	GLHYMPS	(Gleeson et al., 2014; Huscroft et al., 2018)
Porosity	***Por***	-	Discrete*	1 km	GLHYMPS	(Gleeson et al., 2014; Huscroft et al., 2018)
Drainage density	***DrainDens***	1/m	Continuous	250 m (7 arcseconds)	Scheider et al.	(Schneider et al., 2017)
Monthly precipitation	***Precip***	mm/day	Continuous	0.5° x 0.5°	CRU-TS v4	(Harris et al., 2020)
Potential evapotranspiration	***PET***	mm/day	Continuous	0.5° x 0.5°	CRU-TS v4	(Harris et al., 2020)
Diurnal temperature range	***ΔT***	°C	Continuous	0.5° x 0.5°	CRU-TS v4	(Harris et al., 2020)
Land use type	***LandUse***	-	Discrete	1 km	GLC 2000	(Bartholomé & Belward, 2005)
Dominant soil group	***DomSoil***	-	Discrete	1 km	FAO global soil map	(Batjes, 1997)
Water table depth	***WTD***	m	Continuous	Point data	Fan et al. global database	(Fan et al., 2013)

3.2.1.1 Fan et al. database

The database of groundwater table depth built by Fan et al. (2013) has been extensively used in studies since its release in 2013 (de Graaf et al., 2019; Gleeson et al., 2016; Hengl et al., 2017; Maxwell et al., 2015; Pokhrel et al., 2015). The database consists of over 1.5 million measurements of water table depth (WTD) obtained from local, regional, and governmental

monitoring programs supplemented with literature data. However, even though the size of the database is extensive, the data density around the world is highly variable. Figure 3.2 gives the data point density (per grid cell) on a 0.1° spatial resolution. The regions with the highest point density (the Netherlands, Ontario, Canada etc.) are developed countries/regions with relatively high population density and high GDP. The skewness of the dataset towards these parameters is a possible limitation on the usefulness for extrapolation. In addition to the water table depth values, the database also provides surface elevation values for most locations. For full details about the dataset, we refer to the paper itself.

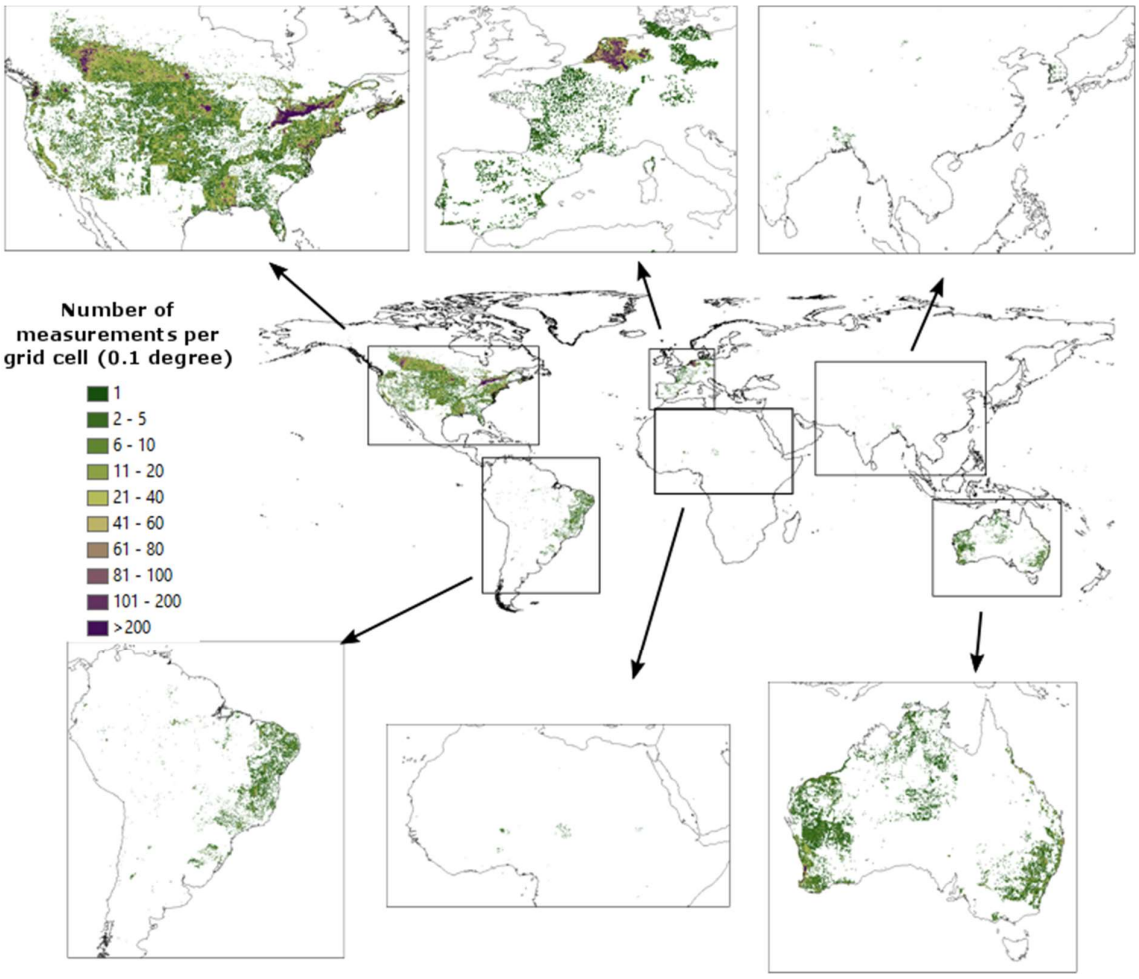

Figure 3.2 Data density of water table depth measurements within the Fan et al database, resampled to 0.1° spatial resolution (Fan et al., 2013).

For this study, we also split up the total (Global) dataset into smaller databases and to extrapolate for these countries/states or regions (Table S-2). These are the individual countries of i) France, ii) Brazil, iii) Australia and iv) the (continental) United States; the individual states of v) Colorado and vi) California and regions across country borders such as vii) the Rhine-Delta region, viii) the Iberian Peninsula. Details about each database can be found in the supplementary information (Appendix B). These smaller databases cover a range of hydro-geo-climatic regions (Reinecke et al., 2020; Winter, 2001) and are used to find

differences in feature importance for establishing water table depth with changing hydrological settings.

3.2.1.2 Climate Research Unit gridded Time Series (Version 4, CRU-TS v4)

For our climate predictor parameters, we used the most recent release of the widely used Climate Research Unit gridded Time Series database (CRU-TS v4) (Harris et al., 2020). This database contains data of monthly observations in ten observed and derived climate parameters obtained from thousands of observation stations in the period of 1910 – 2009. The observations are averaged to monthly values and interpolated to a 0.5° x 0.5° grid to cover regions with no measuring stations. For our study we obtained the Precipitation (*Precip*), Diurnal Temperature Range (*ΔT*), and Potential Evapotranspiration (*PET*) data and calculated the average monthly rates. PET and Precip are proxies for the amount of expected runoff, while ΔT is not in itself a control variable for groundwater table depth but is related to the interannual fluctuations in groundwater table depth and the proximity to ocean water. Precip is directly measured at stations in the dataset, ΔT is calculated as the difference between the maximum annual temperature and minimum annual temperature. PET is derived using the Penman-Monteith equation (Cai et al., 2007).

3.2.1.3 Global Multi-resolution Terrain Elevation Date model (GMTED2010)

To represent land surface elevation in our database we used the Global Multi-resolution Terrain Elevation Date model (GMTED2010) developed by the United States Geological Survey (USGS) (Danielson & Gesch, n.d.), which is a staple dataset to represent land surface elevation in global models. The digital elevation model offers three resolutions (250 m, 500 m and 1000 m) of which we used the latter. The dataset consists of three variables: 1) minimum elevation, 2) maximum elevation and 3) mean elevation of which we used the mean elevation as our data for surface elevation in our model (SurfElev). From this DEM model we calculated the slope of each model cell by dividing the difference between the minimum and maximum elevation by the spatial resolution. This procedure was performed in ArcGis. We chose this dataset mainly, because it was available at multiple (fine) spatial resolutions, and this gave us the initial possibility of making the final groundwater table map as fine as possible.

3.2.1.4 Drainage density

Drainage density was chosen as the proxy parameter for surface water-groundwater interactions. Drainage density is defined as the total length of stream per surface area (m^{-1}). This parameter incorporates aspects of climate and surface water influences on water table depth. We used the global database of drainage density by Schneider et al. (2017) as our source data. The use of this dataset restricted our total predicted area, since this dataset does not support high latitude regions in the northern hemisphere, including Scandinavia, Siberia and the northern provinces and territories of Canada.

3.2.1.5 GLobal HYdrogeology MaPS (GLHYMPS) dataset

We used permeability (Perm) and porosity (Por) as our parameters describing the hydrogeological properties of subsurface. We used the commonly used GLobal HYdrogeology MaPS of permeability and porosity by Gleeson et al. (2014) as the source material of our database. While permeability [m^2] (or hydraulic conductivity [m/day]) and porosity [-] are important for determining groundwater flow in the vertical and lateral direction, for this study we included these parameters since they serve as indicators of infiltration potential. Higher porosity and permeability values indicate a more porous soil/bedrock with higher potential infiltration rates.

3.2.1.6 Global Land Cover 2000 Land Use model

The Fan et al. database is supposed to represent a pre-industrial revolution water table depth with limited human impacts. However, part of the training data originates from regions where groundwater extraction has taken place for decades such as the Central Valley and the High Plains Aquifer in the United States (Gleeson et al., 2012). Land cover (and *LandUse*) has causal relations to water table depth since this serves as a proxy to population density and water use. We used the Global Land Cover (GLC 2000) dataset by Bartholomé and Belward (2005). The dataset distinguished between 21 different land cover classes based on the Land Cover Classification System (LCCS) by the UN Food and Agriculture Organization (Di Gregorio, 2005; Di Gregorio & Jansen, 1998).

3.2.1.7 FAO Global Soil map

Our final data source is for the Dominant Soil type predictor (*DomSoil*). For this we used the FAO global soil map (Batjes, 1997). This dataset consists of 26 major FAO soil type classes with various sub-classes on 1 km spatial resolution. For this study we only considered the 26 major soil types without using the sub-classification. The reason for doing this was to obtain a threshold amount for each class to optimize random forest training.

3.2.2 Data preprocessing: spatial resolution and overlap index (OI)

Since not every predictor data source is available at the right resolution, the question arises whether all predictors are usable at every spatial resolution. Combined with the question whether the water table depth data is usable at fine spatial resolution (<1000m x <1000 m) we initially want to find the optimal spatial resolution for predicting groundwater table depth.

Data preprocessing was performed in ArcGis and involved the following steps:

1. The Fan database was resampled to four different spatial resolutions (0.05°, 0.1°, 0.25° and 0.5°). For high density regions with multiple data points within the same grid cell, we took the mean value of groundwater table depth.
2. The predictor data was also resampled to the same spatial resolutions as for the Fan database. For continuous data we used bilinear interpolation for transforming data

into a coarser spatial resolution. Bilinear interpolation prevents the resampled data to have lower minimum and higher maximum values than the source data. For discrete (class) predictor data we used the most common class during resampling to coarser grid resolution.

3. To represent the global distribution, we transformed the predictor raster data into point data with one data point for every grid cell.
4. To complete the database, we extracted all predictor data at all point locations and exported the data to text files.

To determine the relative representativity for the Fan et al. database to the global distribution in terms of predictor data, we used the novel distribution-free Overlap Index (Pastore and Calcagni, 2019; Pastore, 2018) as our metric for comparison. The Overlap Index (η) represents the percentage overlap between the kernel density distribution functions for two distributions and is defined as:

$$\eta(FGD, GD) = \int min[f_{FGD}(x), f_{GD}(x)]dx \quad (3.1)$$

$$\eta(FGD, GD) = \sum min[f_{FGD}(x), f_{GD}(x)] \quad (3.2)$$

Where $f_{FGD}(x)$and $f_{GD}(x)$are the distribution of predictor x within the Fan et al. database and global distribution (GD) respectively. Equation 3.1 is valid for continuous data, while equation 3.2 is the form of the equation for discrete (class) data. The values for $\eta$are normalized between [0,1], where 1 represents two identical distributions and 0 represents no overlap between the distributions. The reasons for using this metric are simplicity in use and the results serve as proxies for the expected correlation (R^2) for the random forest model. The R package "Overlapping" (Pastore, 2018) was used to calculate $\eta(FGD, GD)$for every predictor at four spatial resolutions (0.05°, 0.1°, 0.25° and 0.5°). The results of this analysis can be found in the supplementary information of this paper (Appendix B, Table B-3). The results show that the minimum $\eta$for all predictors at 0.1° resolution is higher than 0.5. Maximum ηat this spatial resolution is calculated for SurfElev (η= 0.78). Based on these numbers and the desire to produce a map on the finest spatial resolution possible, we decided to use 0.1° as our target for our water table depth map. The full schematic process for data preprocessing is found in Figure 3.3.

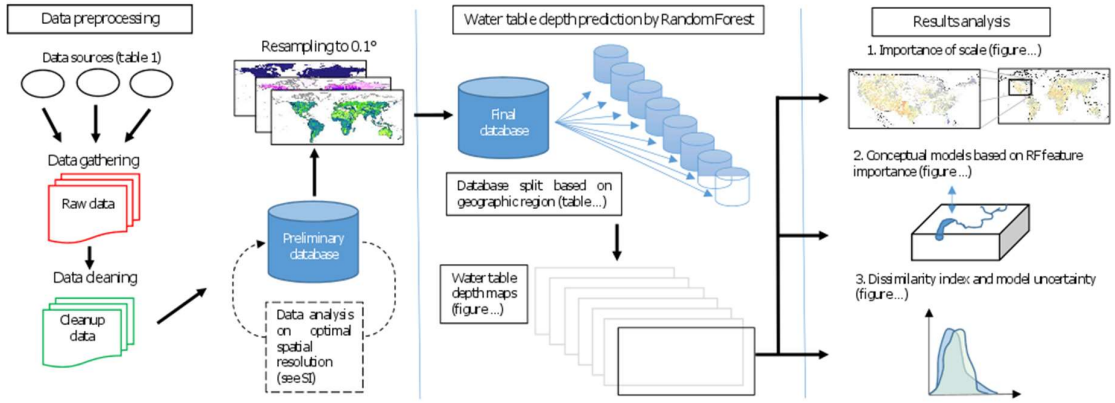

Figure 3.3 Workflow, references to tables and figures and deliverables of this study indicating the three main contributions to the scientific understanding of machine learning models predicting water table depth.

3.2.3 Random forest and feature importance

A random forest model is a widely used machine learning algorithm capable of regression and classification tasks (Liaw & Wiener, 2002; Naghibi et al., 2016; Prasad et al., 2020). A random forest is an ensemble method of where an *N* amount of decision trees are regenerated based on the test data. The random forest gives the average prediction over all generated Decision Trees. The random forest model is useful since the potential overfitting on the training data is negated when the number of decision trees is high. The tradeoff for this can be potentially comprehensive learning time when the dataset is large. Unlike with artificial neural networks (Govindaraju & Rao, 2013; Sreekanth et al., 2009), random forests do not require feature scaling and can both handle numerical (continuous) data as well as categorical (non-numerical) features simultaneously.

To determine our model performance, we applied 10-fold cross validation for each database individually by randomly splitting the total database between 70% of training data and 30% of cross validation data. The 10-fold cross validation was used to check whether random splitting of the data would give high variance in model accuracy. The variance between each model was found to be very small (variance in R^2 of ± 0.03).

For each model we calculated the cross-validation error using three different scoring metrics: 1) the coefficient of determination (Pearson R squared, R^2), 2) the mean absolute error (*MAE*, equation 3) and 3) the root mean squared error (*RMSE*, equation 4). Here y_{pred} and y_{obs} are respectively the predicted and observed variables.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_{pred} - y_{obs}\right| \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{pred} - y_{obs}\right)^2} \quad (4)$$

Feature importance was determined by two different methods which showed similar results: for the first method we used the `feature_importance` function in the `Scikit-learn` toolkit. Using the `feature_importance` function we calculated the fraction of total splits within all decision trees which are determined by each predictor. For discrete parameters which were split by one-hot encoding, the total feature importance of that predictor was based on the sum of the feature importances for all splits. The second method for determining feature importance was to replace predictor data with randomized one predictor at a time. Then we observed the change in model performance between our base run and the run with randomized data. The most important features would cause the largest drop in model performance. For predictors with continuous data, we replaced the randomized data following a Gaussian distribution with a mean of $\underline{x}$ and a standard deviation of $\frac{\underline{x}}{3}$, where $\underline{x}$ is the mean of the predictor over the entire training data. Choosing this particular standard deviation ensures that around 98% of data points generated will fall between the minimum and maximum value for that predictor in the training set. Discrete predictor data was replaced with a discrete uniform distribution among all possible classes. Both methods for determining feature importance are essentially methods for sensitivity analysis, of which the importance has been made clear by Razavi et al. (2021).

To produce the final water table depth maps, we differentiated between areas where water table depth was already known based on the training data from the Fan et al. database (from now on called the *training domain*) and areas where water table depth was unknown and is extrapolated to (from now on called the *predicted domain*). An example of a finalized water table depth map for the USA is given in Figure 3.4.

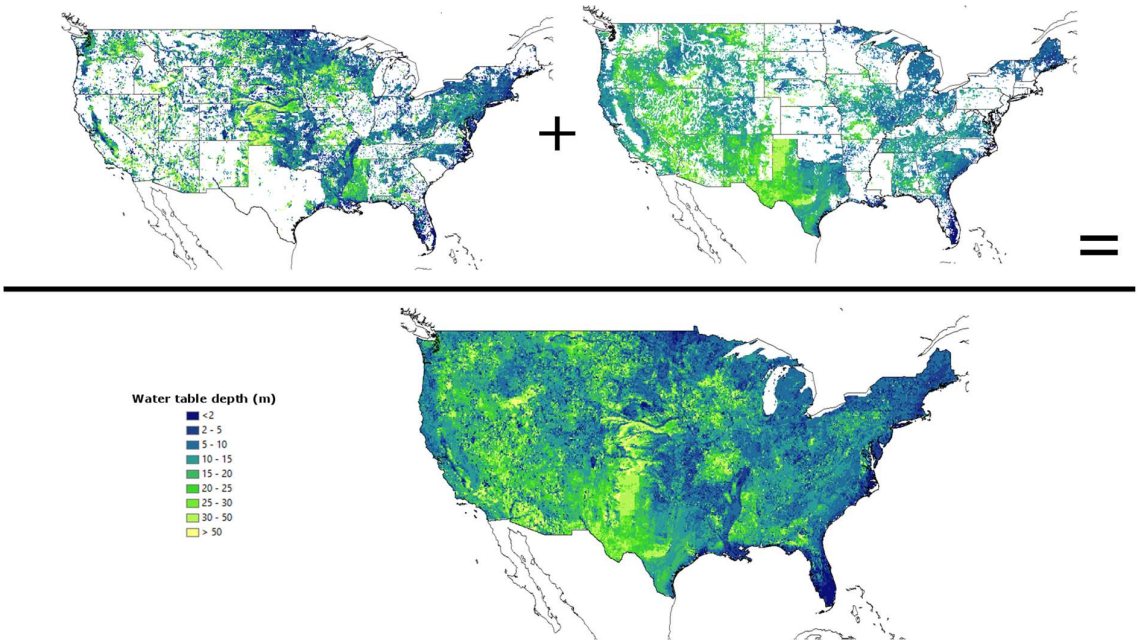

Figure 3.4 Example of the building of the final water table depth maps in this study as the sum of areas of known data (*training domain (top left))* and the regions of unknown water table depth where we extrapolated our model to *(predicted domain (top right))*. The combination gives the final water table depth map (bottom). The full maps for all regions are found in the supplementary information (Appendix B).

3.2.4 Dissimilarity Index (DI)

Our final analysis on the usability of our water table depth map is to determine the level of confidence we have in the extrapolated results. To quantify this confidence, we calculated the Dissimilarity Index (DI) for every predicted model cell outside the training domain. The DI is a metric of similarity between a new target cell relative to all the data cells within the training set within *N*-parameter space, where *N* is the number of predictors. This method is adopted from Meyer and Podesma (2020) and applied to this study.

The DI is defined as Euclidean distance of test sample in *N*-parameter space to the nearest training data point d_k, divided by the mean Euclidean distance between every two-point combination within the training dataset ($\underline{d}$) (within the same *N*-parameter space). We calculated the DI as is described by Meyer and Podesma (2020). Prior to calculation all predictor parameters were scaled by calculating the Z-score to remove any unit dependencies. A scaled input parameter X for sample $i$is calculated by:

$$X_i^S = (X_i - \underline{X})/\sigma_X \tag{3.5}$$

Where X_i^S is the scaled parameter value for sample i of predictor X, $X_i$is the unscaled parameter value, $\underline{X}$ is the mean value of predictor X over all samples in the training set and $\sigma_X$is the standard deviation of predictor X over all samples in the training set.

Furthermore, we applied a weighting to the scaled samples based on the feature importance in the random forests model. The parameters were scaled based on the feature importance in Figure 3.4a:

$$X_i^{s,w} = w_X X_i^s \tag{3.6}$$

Where $X_i^{s,w}$is the weighted and scaled predictor value for sample i and $w_X$is the feature importance of predictor X. From here we can calculate $d_k$finding the minimum Euclidean distance between each test sample k and the training set.

$$d_k = arg\ min\ d(k,i) \tag{3.7}$$

The Euclidean distance between two points $p$and $q$within N-parameter space is given by:

$$d(p,q) = \sqrt{\sum_{j=1}^{n}(X_{p,j}^{s,w} - X_{q,j}^{s,w})^2} \tag{3.8}$$

3.3 Results and discussion

3.3.1 Random forest results and feature importance

Table 3 gives the results of the trained random forests models for each individual database. The scatter plots and maps of the cross-validation results for each individual database are given in Figure 3.5. The highest correlation was observed for the Colorado and Australia databases (R^2 = 0.794 and R^2 = 0.857 respectively). For the global water table depth map (Figure 3.5a) the random forest model has a correlation of R^2 = 0.717 (MAE = 4.584 and RMSE = 8.956) to the training data. The weakest model correlations were found in the France and Iberia models with relatively the highest MAE and RMSE values and lowest correlation. Model performance is not correlated with the fraction of total area covered by training data (training domain) or database size. For example, two models (Iberia and Australia) with almost identical coverage by training data have the lowest and highest correlation. These percentages are normalized by difference in total surface area. Training domain coverage is calculated by dividing the number of training samples over the total number of grid cells within the total domain. From here we deduce that feature importance is more important than database size and that our models were not underfitted due to lack of data. Figure 3.7 shows the combined global map of water table depth generated by the random forest model. The maps of the other databases can be found in the supplementary information (Appendix B).

When looking at the cross-validation results in Figure 3.5 and Table 3.3, we can conclude that all models are slightly underestimating regions with higher water table depth relative to the training data. We attribute this to the fact that the Fan et al. database is underrepresented in

samples at higher elevations and remote drier climates and overrepresented in more densely populated moderate climates and areas with relatively higher Gross Domestic Product. Since the database has most data from easily accessible, densely populated regions rather than sparsely populated remote/mountainous regions, we expected a skew of the final model towards those densely populated regions. Even in the smaller, low relief Rhine Delta model, higher water table depths are underestimated. A second reason comes from the random forest algorithm itself. Due to the lack of data of high water table, it is harder on the model to create distinguishable leaf nodes for higher water table depths. This parameter is manageable with the minimum samples per leaf node hyper parameter (`min_samples_leaf`), which we already tuned to be very low (2-4 samples per leaf node). The consequences of underrepresenting water table depth is further discussed in section 3.4 of this paper.

The results of our feature importance analysis are given in Figure 3.6. The first metric was used on all datasets (Table 3.3), while the second method was only applied to the total (Global) database. Figure 3.6 gives the feature importance for our Global model, with the error bars indicating the standard deviation of the particular predictor over all over models. The water table depth maps and bar charts with feature importances of individual databases are added in the supplementary information (Appendix B, Figures B-4 – B-16). Based on the first method we found *SurfElev, Slope* and *DrainDens* are considered to be the most important predictors for our model, while *DomSoil, LandUse* and *Por* were considered least important. The results from method 2 (replacement with randomly generated data) shows *SurfElev, Precip, DrainDens* and *DomSoil* to be the most important predictor parameters since changing these predictors caused the largest drops in model performance (R^2). *Por*, *Slope* and ΔT were found to be least important using method 2. Since all drops in model performance were significant (> 0.1), we argue that all parameters were important for our model, and we haven't used unrelated predictor data. We did not perform an extensive analysis that proved all predictors are statistically independent. For example, one might argue that *Precip* and *DrainDens* are related parameters since areas with high annual precipitation also have a denser stream network. (Dingman, 1978; Tarboton et al., 1992). However, we did find that feature importance varied among different databases (Appendix B, Figures B-4 - B-16).

Interestingly, according to the results of the first method, the random forest model seems to favor continuous data over discrete data since the predictors consisting of discrete data were almost all ranked least important. This conclusion cannot be drawn based on the results of the second method. With our current understanding of hydrological knowledge, we would expect surface elevation, drainage density and precipitation to be factors important for steady-state water table depth. Porosity and permeability are important parameters for determining groundwater flow vertically and laterally. We explain this to the fact that contrary to many recent studies in this study we predict a steady-state water table depth rather than a transient water table depth or hydraulic head (Bowes et al., 2019; More, 2018;

Zhao et al., 2020). Since porosity and permeability are indicators of groundwater mobility, we would expect a higher feature importance if we would have predicted transient water table depth. Also, we argue that part of the discrepancies between the results of the two methods lies in the methodology itself combined with the skewness of the data. Some predictor data (*LandUse* and *DomSoil*) are heavily skewed towards only a few classes where most of the samples fall in. Some LandUse classes and DomSoil groups only have a few samples within a few classes. With our second method of determining feature importance, we used a uniform probability distribution between all possible classes. The randomly generated surrogate data would therefore be less skewed than our original data.

Table 3.2 Database characteristics and model correlation (validation errors) for trained random forest models.

Database	**Database characteristics**			**Model correlation (validation error)**		
	Size (# of samples)	**% coverage training domain**	**% coverage predicted domain**	**R^2**	**MAE (m)**	**RMSE (m)**
Rhine Delta	954	100	0	0.734	1.391	2.535
France	1686	26.34	73.66	0.639	6.761	11.349
Iberia	766	12.64	87.36	0.681	8.767	12.887
USA	39863	48.72	51.28	0.711	4.824	7.507
California	1276	31.06	68.94	0.732	5.700	8.975
Colorado	1456	51.72	48.28	0.794	5.493	8.548
Brazil	8320	11.83	88.17	0.690	4.770	7.920
Australia	10774	15.68	84.32	0.857	2.394	4.396
Global	79880	6.96	93.04	0.717	4.584	8.956

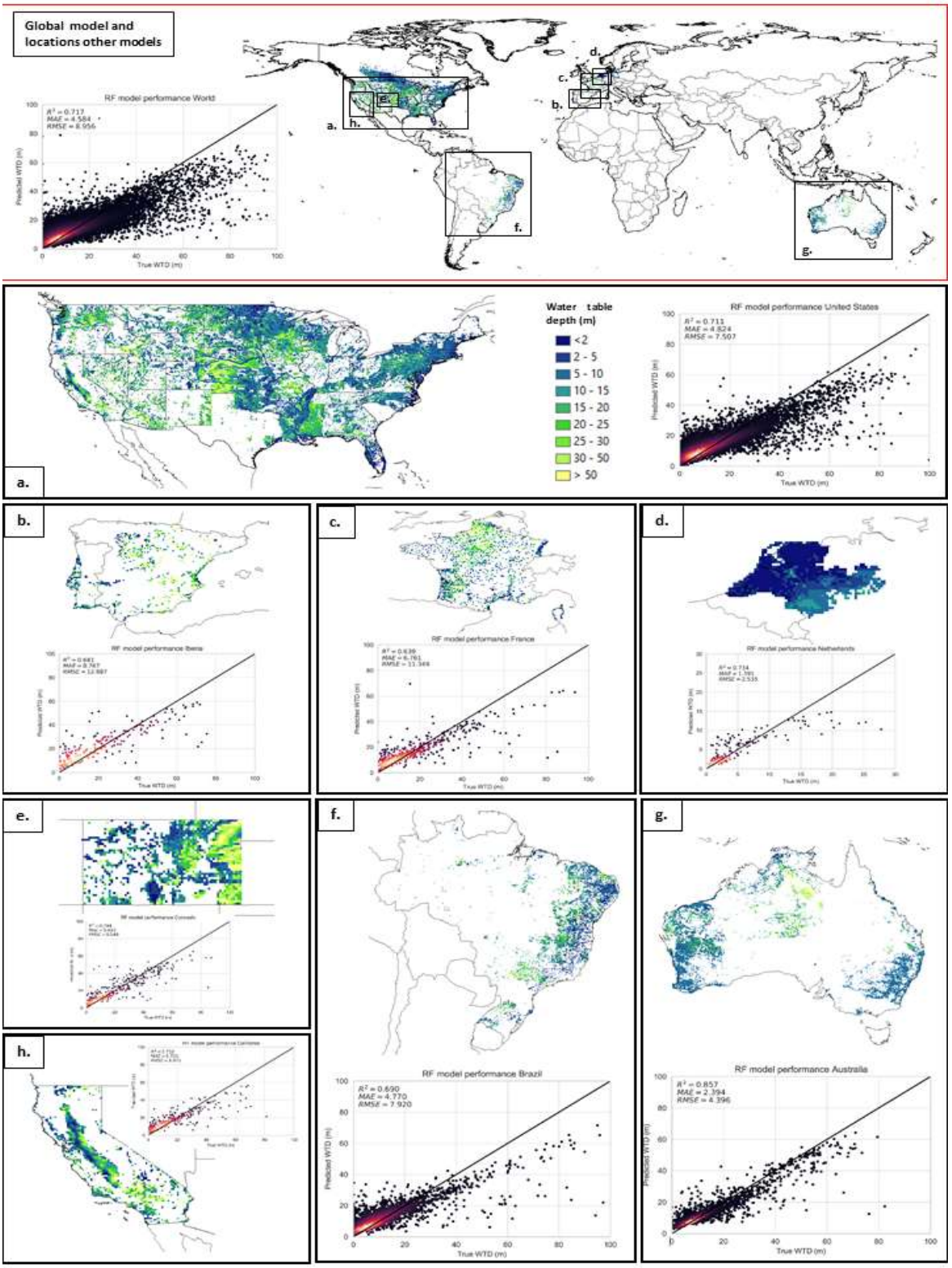

Figure 3.5 Random forest model results for water table depth for the Global model (top) and every sub database: a) United States (excluding Alaska, Hawaii and overseas territories), b) Iberia, c) France, d) Rhine Delta, e) Colorado, f) Brazil, g) Australia and h) California. The scatter plots show the cross-validation errors and correlations (mean absolute error (MAE), root mean squared error (RMSE) and correlation (R^2) for every model. The diagonal represents the 1:1 correlation line.

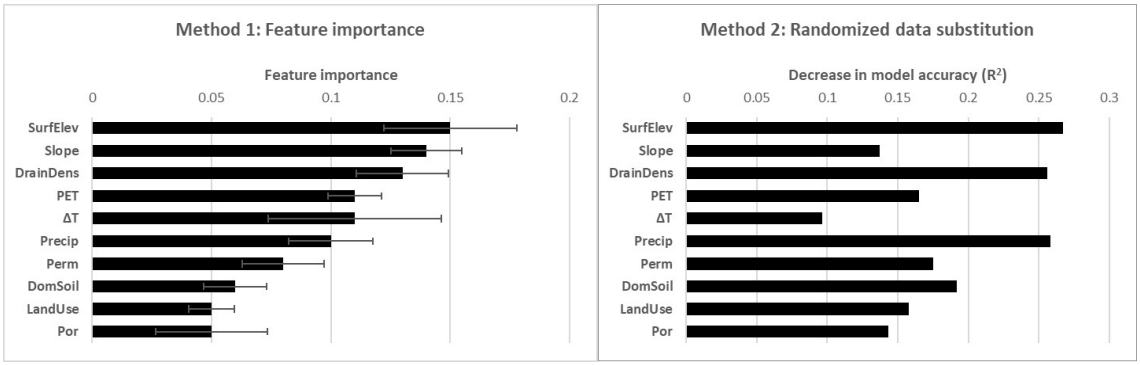

Figure 3.6 Results of feature importance analysis. Left: Ranked feature importance as calculated by Scitkit-learn's `feature_importance` function. The horizontal bars show the feature importance within the Global database. The error bars show the standard deviation of each individual predictor over every other database. Right: reduction in feature importance as predictor is replaced by randomized data.

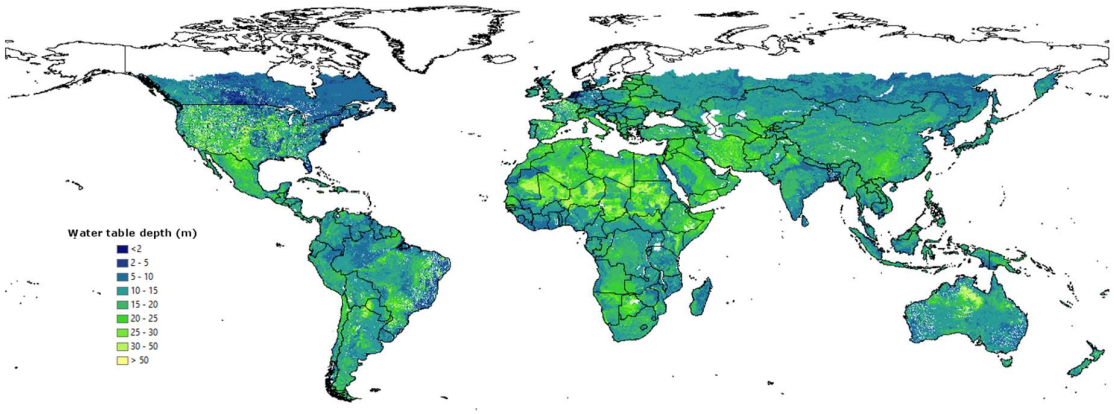

Figure 3.7 Global map of water table depth (m) based on the random forest model (for the predicted domain) combined with the training data from the Fan et al database on a 0.1-degree spatial resolution.

3.3.2 Model performance compared to physics-based models

We can check the extrapolated water table depth of our model to other global groundwater models which estimate hydraulic head or water table depth. Even though this does not give any direct truth on the validity or accuracy of our random forests model, this does serve as a proxy on how our model performs against a numerical model based on the known physical laws and controls of groundwater (flow). For this model performance analysis, we chose three different numerical models: 1) the numerical model for water table depth from Fan et al. (2013) which is based on our training data, 2) the high-resolution groundwater De Graaf model (De Graaf et al., 2015; Sutanudjaja et al., 2018) and 3) the G³M model (Reinecke et al., 2019; Reinecke et al., 2019). Figure 3.8 shows the correlation matrix for water table depth between our random forest model and the three physics-based models. The data is represented with heat maps for each 1 m of water table depth. Darker shades indicate higher point density for that water table depth. The gridded data from the physics-based models was resampled to the same 0.1° grid resolution as our random forest model (using bilinear interpolation as our resampling method) in order to make a fair comparison.

The first column in Figure 3.8 shows the correlation between our random forest model and the three physics-based models. The physics-based models show higher values for water table depth compared to the random forest model. The Fan et al model has the best correlation to our model. This is expected since both models are calibrated to the same data source. We argue that there are few explanations for the discrepancies between the different model types. 1) One of the limitations of the random forest model compared to physics-based models is that there is a limited range of values the model is able to predict for. Our random forest model is only able to predict values in the range of the leaf nodes with the lowest and highest values for water table depth. 2) Our random forest model has a drastically different model setup, compared to the other models. The main difference between our model and the three other models is that the three physics-based models are more or less based on the same theorems of physics, with some discrepancies. Interestingly, when comparing the results of the three numerical models to each other directly we also find no real significant correlation, except for G^3M to De Graaf (Figure 3.8, bottom row, 2nd from the left). This is contributed to the fact that contrary to the models of G^3M and De Graaf, the Fan et al model does not simulate surface water interactions. The relative disagreement between numerical models is here found to be comparable to the relative disagreement between a physics-based model and a machine learning model. From here we conclude that our model does not perform better or worse than high-end numerical models with some notable differences/limitations: 1) The random forest model is only able to simulate water table depths within the available data range (< 56 m below ground surface). This shallower water table depth is more realistic than water table depths simulated by the physics-based models, which go up to several kilometers for the PCRGLOB-WB model. 2) Our simulation of water table depth was performed on a 0.1° spatial resolution (~ 10 km). To adjust for a relatively coarse resolution, we had to calculate the mean water table depth among all water table depth measurements for each grid cell. Since some areas, such as Ontario, Canada have a high measurement density we risk losing important information when averaging out water table depth. We argue that we lose the representation of this high variance in water table depth when we reduce >100 measurements to a single datapoint per grid cell as has been shown by Reinecke et al. (2020).

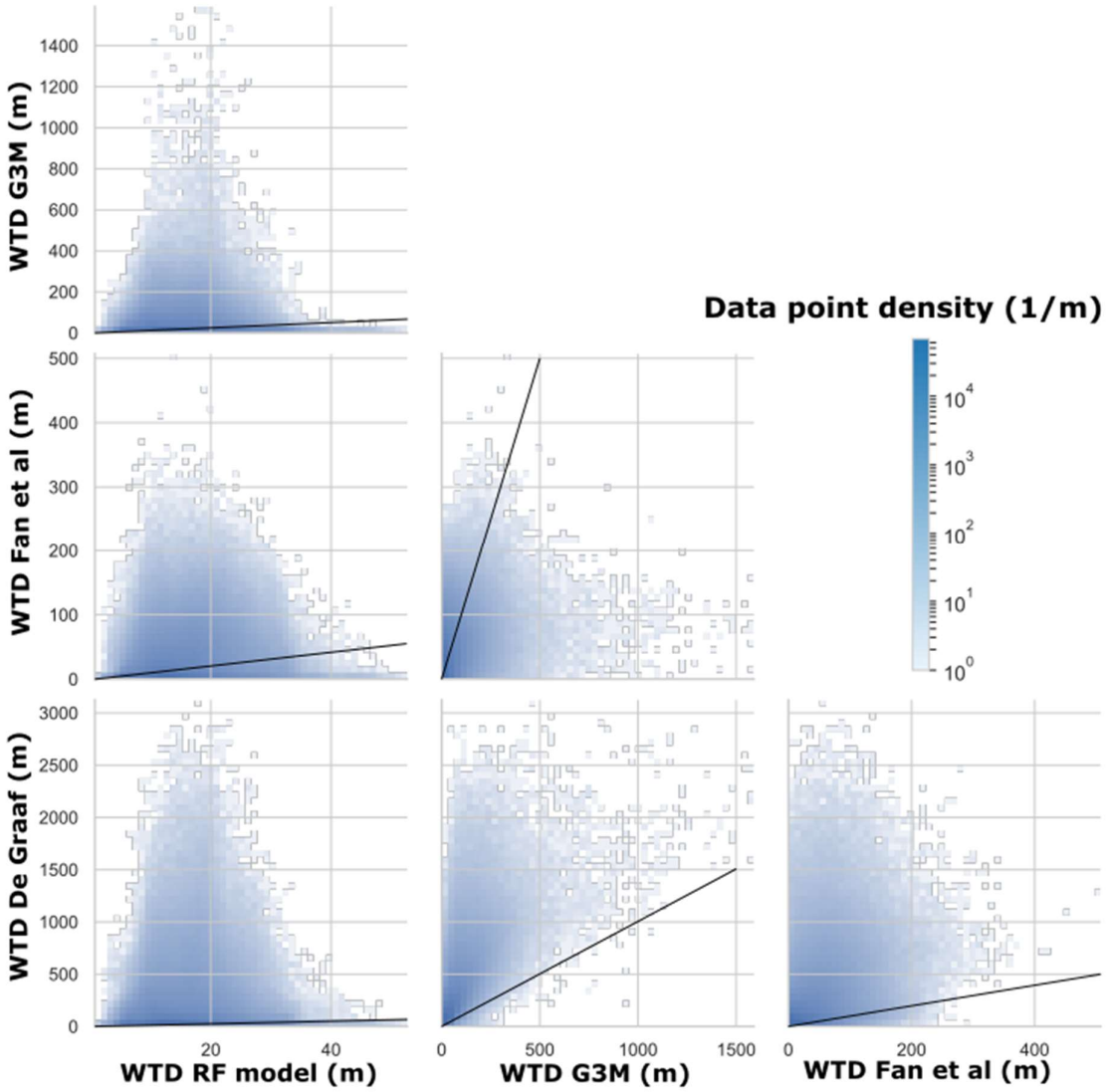

Figure 3.8 Pair grid plot of model correlation between four distinct models: 1) The random forest model for water table depth (this study), 2) Fan et al model (Fan et al., 2013), 3) G^{3}M model (Reinecke et al., 2019) and 4) De Graaf model (De Graaf et al., 2015). The figures show heatmaps of water table depth measurements (darker shade indicates higher point density) and the solid lines depicts 1:1 lines for WTD.

3.3.3 Feature importance and perceptual models from our random forest model

Figure 3.1 showed schematizations of the similarities and differences between numerical model studies and machine learning studies on the same topic and how the underlying study fits within that machine learning organization (Figure 3.9). We found significant differences in feature importance between a Global model (using all training data) and our Rhine Delta model (using limited data from that particular region (Figure 3.9a). These databases represent two quite different hydro-geo-climatic environments: 1) an area with Low Relief (Rhine Delta) and 2) an area with High and Low Relief (Global). While the relatively simple Rhine Delta region can be described by 57% based on just three predictor parameters (*SurfElev*, *Slope* and

DrainDens), these same predictors together only explain the Global model by 42%. Several reasons could be given for this. First of all, the Rhine Delta region is extremely small compared to the Global model with relatively uniform climate, low relief and low variation in land use. The random forest model takes the three parameters which have highest variance and defines these as the defining parameters to split the nodes within the random forest. On the other hand, the Global model is on the other end of the spectrum in terms of complexity. The Global model is able to predict for any regions of the world which are in the same domain as the predictor data of the training set. Even though not all climate regions or elevations are equally represented, this still covers a large domain of the global map. To define water table depth within this larger range of hydro-geoclimatic regions, more predictor parameters are needed. Consequently, a more spreaded feature importance is observed by the random forest model.

A hydrologist with limited knowledge on the workings of machine learning algorithms might wonder whether these results can then also have some physical meaning. One part of the physical relevance of the model comes from the choice of predictor parameters. As can be seen in Figure 3.9, the choice of predictor parameters is very important for a robust machine learning model. Choosing too few parameters might result in an under-fitted model which lacks capabilities of predicting the target variable. Choosing too many variables gives the risk of overfitting the model to the possible noise in the data and those different predictors might give redundant information, causing further noise. From a hydrologists point of view the Low Relief model seems to be too simplistic to give a thorough representation of the different parameters determining groundwater table. The scale of the regions comes into play here. The Rhine Delta region spans only 400 kms laterally compared to the larger Global model. On this scale parameter variance of climate predictors (*Precip*, *PET* and *ΔT*) are much smaller since the lateral differences in climate are small. Therefore, we see the highest feature importance in non-climatic geographical parameters (*Slope*, *SurfElev* and *DrainDens*). From here we find scale to be an important factor when deriving a perceptual model from random forest results: we cannot find a single perceptual model for the entire world but find different perceptual models for different scales/continents. Future research could include building a water table depth map for the Rhine Delta region on a much finer spatial resolution. In this way predictor variance should increase over the relatively coarse grid resolution of the current study. Predictor variance will increase this way too, which in turn might increase the feature importance of climate predictors such as PET and the perceptual model might become more similar to the perceptual model of the Global model. It has already been proven that it is possible to make a model for predicting water table depth for the Rhine Delta region using only global parameters (Sutanudjaja et al., 2011). Sutanudjaja et al. used global data to calibrate a regional groundwater model for the Rhine-Meuse basin in Germany.

One other important difference between a physics-based groundwater model and the current study is the lack of connectivity between adjacent model cells. Rather, all cells in our random forest model are predicted independently based on the predictor data for that cell only.

Unlike numerical (MODFLOW) models there is no-intercell connectivity which determines the hydraulic head field. However, even without this dependence of the target cell to its adjacent cells, the random forest model results still show continuity across regions and on the boundaries between regions with training data and predicted regions. This is clearly shown in the map of the United States (Figure 3.4). While the water table depth in the northern part of the High Plains Aquifer (Nebraska, Kansas, etc.) is known from the Fan et al. database, we observed a continuity of the deeper water table depths into the predicted regions of Texas. Additionally, the shallow water table depths representing swamps and wetlands in Florida are also observed. This is a positive result for a model with no interconnectivity between model cells. Therefore, the random forest model is able to simulate a lateral water table gradient based solely on the parameter gradients of the input parameters. Other features we discovered is the ability of the random forest model to simulate a shallower water table depth next to ocean waters and large surface water bodies. This can also be observed from the map for the USA random forest model (Figure 3.3), where the states of Georgia and South Carolina show continuity between the known data regions of Florida and North Carolina.

How is the modelling approach shown in this study beneficial to the larger scope of modelling studies involving groundwater? We see reasons to prefer this approach over the conventional approach outlined by Hill and reviewed by Zhou (Hill, 2000; Hill & Tiedeman, 2006; Zhou & Li, 2011). Zhou reviewed the methodology of creating regional scale groundwater models and stated that a groundwater model for the purpose of resource assessment should be built at the basin scale and should include 1) topography, 2) major regional aquifers and aquitards and 3) groundwater-surface water interaction since they are essentially the same (interconnected) source. The spatial resolution of a numerical model can range from a single to several kilometers. With this comes added difficulty on how to scale feature parameters properly. While hydraulic conductivity can be measured at point locations (individual wells or measurement locations), upscaling the parameter to a larger scale can severely limit its accuracy and usefulness. Sobieraj et al. (2004) concluded from their field study on variations in hydraulic conductivity, that there is little predictability for hydraulic conductivity on the field scale due to local biological processes controlling soil hydraulic conductivity. Gupta et al. (2006) came to a similar conclusion based on their own field test.

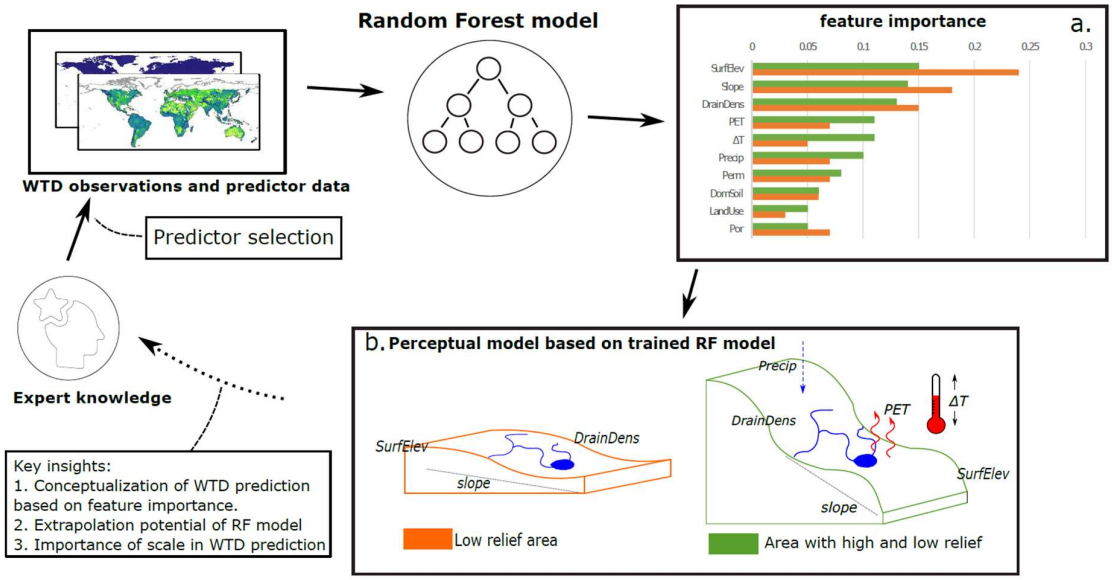

Figure 3.9 Representation of this study based on the schematic representation of geosciences within a machine learning framework (Figure 3.1b). Subfigures: a) the feature importance for the Global database (green) and the feature importance for the Rhine Delta database. b) Translation of the feature importance from to two conceptual models based on the results of the random forest model for an area with low relief (orange) and for an area with high and low relief (green).

3.3.4 Importance of scale

Model scaling has always been a difficult problem for hydrological models (Blöschl, 2001; Blöschl & Sivapalan, 1995; Fraser et al., 2013; Gentine et al., 2012; Peters-Lidard et al., 2017; Scheibe & Yabusaki, 1998). While Darcy's law and the groundwater equation might suffice for local or regional scale models, the distribution of water table depths is also controlled by topography, geology and climate and data for these at continental to global scales can be challenging. A difficulty in this for hydrologists is how to transfer these hydrological processes to the larger scale and to build the numerical model accordingly. Most of the thought process has to be done upfront and reflection only begins until after the model has been built and the results are generated (Figure 3.1a). For machine learning studies we start building our model based on some level of expert knowledge and available data. From there we improve our model by limiting underfitting (for example by adding more features or obtaining more training data) and limiting overfitting (for example by removing redundant features). We looked at this problem through two different lenses: 1) how does a large-scale model (for example the United States as a whole) perform on a small subdomain compared to a model that is only trained on this subdomain and 2) how does feature importance within random forest models change across scales? For the first question we looked at three of our random forest models (global, USA and California) to see if larger scale models were able to determine water table depth as accurately as their smaller scale counterparts. The second question is tackled in section 3.5. Figure 3.10 shows the water table maps of these three models and the correlation of scatter plots between these maps. Looking at the differences in the maps for

California, we observe that the smaller scale maps generally show shallow groundwater tables compared to the large-scale models. This is partly explained by differences in feature importance (section 3.5, Figure 3.11), but also due to the size and composition of the database. When building a larger scale model which encompasses a larger range of geoclimatic regions we must account for this larger range with more variance in input data. Some of the added data will represent areas which are geoclimatically different from the climate and hydrogeology of California. When comparing the maps for California derived from the USA and California database respectively, we do observe a decently good correlation (R^2 = 0.705) showing some potential for large scale models here. Feature importance between these models is also comparable (Figures B-4 through B-16).

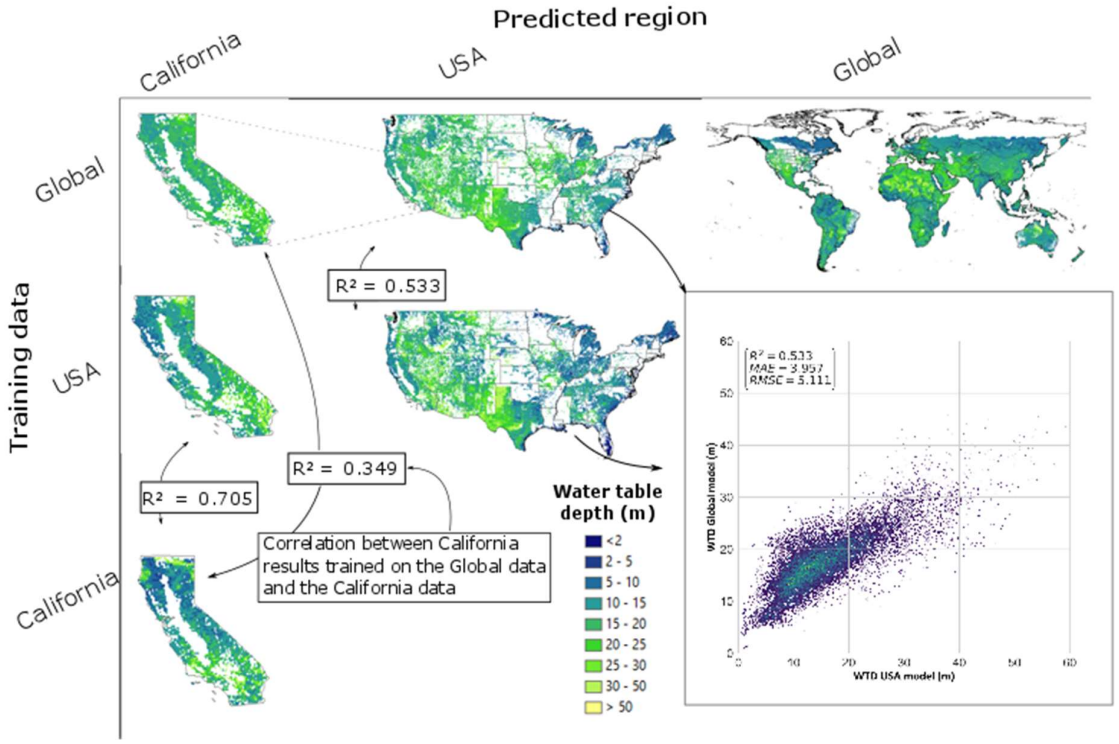

Figure 3.10 Importance of scale when predicting water table depth using random forests based on different training datasets (y-axis) and prediction domains (x-axis). Correlation values between maps are given in boxes. Example of correlation plot is given for two maps.

3.3.5 Extrapolation potential

The Dissimilarity Index (DI) is an indicator of extrapolation potential for our random forest model. The DI represents the similarity between the predicted domain and our training domain within the multidimensional parameter space based on our chosen predictor data. We calculated the DI for the predicted domain of the Global random forest model. The results are represented in Figure 3.11. Here we combined the DI results with the simulated water table depth of our random forest model, where the color range simulates different water table depths and transparency depicts DI (high transparency indicates high DI). The DI results are also given in the supplementary information. Values for DI > 2 were considered outlier data since over 99% of the data points show a DI < 2 (see supplementary information,

Appendix B). Based on the maps we can identify the areas that are statistically most similar to our training set and therefore we expect our model to perform best. These areas include most of Canada, Northern-India among other regions. Areas with the lowest DI are mainly tropical regions around the equator (Amazones, Central-Africa and Indonesia) and high elevation regions such as the Himalayas. The Uttar Pradesh Province in India is one of the regions with the highest potential for our random forest model, given that we have almost no data from Asia in our database. We have several explanations for these DI results.

1. First of all, higher elevation regions are underrepresented in the Fan et al. database and consequently, we have seen that the random forest model underestimates higher values of water table depth. Moreover, given that the DI is weighted based on feature importance and surface elevation is found to be one of the most important predictors for steady-state water table depth. This result is carried through in the calculation of DI and henceforth we see these low DI numbers for the Himalayas and Andes.
2. The same reasoning could be applied for the low DI numbers for tropical climate regions in the Amazon and Indonesia. These DI numbers are arguably primarily driven by high values for precipitation and evapotranspiration, which are relatively under sampled in the training data.
3. The final, relatively surprising, finding is that we see relatively high DI numbers in arid regions such as the Sahara Desert and Middle East. While these regions are almost exclusively predicted and the Fan et al database doesn't include any training data, we speculate the model is still moderately efficient in predicting these regions, given that we do have decent amount of training data from arid regions in the United States and central Australia, which would arguably be decent proxies for estimating arid regions globally.

A final important point we want to emphasize is that the DI is not a metric for estimating *model accuracy*, but rather for *model extrapolation potential*. We argue that the DI is a suitable metric to give meaning to raw model results. While most current studies do not include *hydrologic dragons* ("regions where the uncertainty in expected hydrologic behavior or relevant system properties is very high due to a lack of local knowledge", cited from Wagener et al. (2021)), in this study we have attempted to include this by using this metric. From here we determined for which areas we have most confidence in our random forest model. Future work might include obtaining more data from high DI regions (such the Uttar Pradesh province) in India to test our model against.

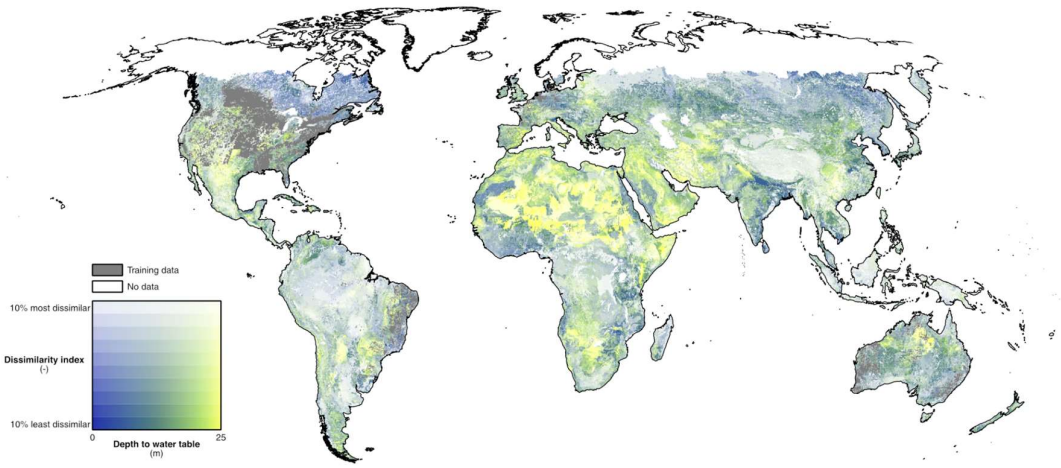

Figure 3.11 Global map of water table depth and dissimilarity Index calculated for the predicted domain on a 0.1° scale. Methodology based on the studies by Meyer and Pebesma (2020). Outlier values for DI (>2) are excluded from this graph. Color range shows water table depth (m) and transparency shows dissimilarity index.

3.4 Conclusions

In this study we constructed a steady-state, global water table depth map on a 0.1° spatial resolution using the random forests algorithm. For the training data we used the well-known (and only) global water table depth database by Fan et al. and we chose hydro-geo-climatic predictors with an assumed causal relation to steady-state water table depth, including surface elevation and drainage density.

The resulting water table depth maps show a correlation of R^2 = 0.7 - 0.85 (cross-validation error) depending on the training domain. With this study we contributed to the scientific knowledge within this field in five meaningful ways:

1. We developed the first global steady-state water table depth map using a random forest algorithm on a 0.1-degree spatial resolution.
2. Based on the results on our random forest model we constructed two perceptual models for high and low relief domains. Above ground variables (surface elevations and climate parameters) are found to be more important than sub-surface parameters (permeability) in explaining water table depth.
3. We observed differences when predicting the water table depth of regions across spatial scales using the random forest models for California, USA, and the Global model. Shallower water table depths are observed when using more small-scale databases.
4. Comparing the distribution of water table depth of the new machine learning model and existing physics-based global hydrologic models highlight significant differences both between models and models and observations.

5. The Dissimilarity Index has shown to be a good initial attempt at quantifying extrapolation potential for machine learning algorithms.

Within this study we hope to contribute to the overall implementation of machine learning algorithms as valuable tools within the field of hydrology. Not only as a pure modeling tool, but to gain insight in the physical mechanisms of (ground)water. We believe this can be an asset for future studies on water availability, water flow and water security.

Chapter 4: Contributions of this thesis to the unsolved problems in hydrology

The goal of this thesis is to contribute to the increasing use of machine learning algorithms in addressing groundwater problems in the hydrological sciences. Based on the literature review in chapter 1 and the findings in chapters 2 and 3, it can be concluded that this research has advanced our knowledge by contributing to three of the unanswered questions in the hydrological sciences (Blöschl et al., 2019). The focus of this study was on the contributions of machine learning in hydrology and not to contribute to the advances of machine learning themselves. In this concluding chapter we return to the three unanswered questions in hydrology mentioned in chapter 1, section 1.4 and reflect on how the results of this Master's thesis have progressed our understanding on these three topics.

1. **What are the hydrologic laws at the catchment scale and how do they change with scale?**

Chapter 3 of this thesis used a random forest model trained on a water table depth database to predict steady state water table depth on the regional and global scale. From the random forest model, we derived the most important feature parameters for each individual region. One of the regions that was predicted is the Rhine Delta region/catchment of Western-Germany and the Netherlands. The results show that the steady-state water table depth is dependent for over 50% based on three feature parameters alone (surface elevation, slope, and drainage density). When we compare this to country sized databases, such as the United States or globally, we find that these three parameters are still among the most important, together with precipitation, evapotranspiration, and diurnal temperature range. This final feature parameter is non-intuitive but can either be interpreted as a proxy for proximity to sea level or as a proxy for temperate climate regions where water table fluctuations are expected to be lowest throughout the year. Surprisingly, geological parameters such as permeability and porosity were found to be among the least important parameters. We contribute this to the idea that these features would be more important when predicting transient water table depth, rather than steady state, where regional groundwater flows and seasonal feature variability is more important. We did not find significant variation in the importance of permeability or porosity across different databases of different sizes. From here we conclude that for this specific spatial resolution geological parameters do not follow a clear scalable relationship. If we place these findings within the context of hydrologic laws, we can state that this research only contributes moderately in terms of hydrologic scaling laws. However, in terms of climatological hydrologic laws this research contributes with the finding that although we observe variances in feature importance for estimating water table depth across different hydrogeoclimatic regions (Appendix B), surface water features are consistently found to be the most important parameters regardless of climate.

2. **How can we use innovative technologies to measure surface and subsurface properties, states, and fluxes at a range of spatial and temporal scales?**

Random forests and artificial neural networks are not "innovative" themselves since they have been used since the 1980s. However, our methodologies can be considered novel in their approach. In Chapter 2, we predicted groundwater and surface water depletion caused by groundwater pumping using an artificial neural network. For our training data, we used synthetically produced data derived from numerical models. The synthetic data was generated based on physical principles of mass balance using a *local area* approach. The local area was based on a mass balance law that stated that all pumped groundwater should be coming either from storage or reduced baseflow. Another novel feature of the artificial neural network is that the neural network attempted to predict depletion at five different output times simultaneously, rather than predicting depletion after a certain pumping time t. This is an example of a multiple response variable model or multi-task model, which have been named recently as an avenue of innovation for machine learning algorithms (Reichstein et al., 2019). Although the artificial neural network has not outperformed numerical models or comparable machine learning studies, this is still a novel attempt. Methodologies that don't work perfectly can still give valuable insight and be good starting points and learning experiences for future work.

3. **How can hydrological models be adapted to be able to extrapolate to changing conditions?**

Finally, both chapters 2 and 3 touch upon the common problem of extrapolation with numerical models and machine learning models. Numerical models and machine learning models often struggle with extrapolation to other areas, because for example they are built and calibrated for a specific watershed (for numerical models) or the training data used to train the machine learning model is too area/climate/region specific.

Chapter 2 tackled the extrapolation problem using a cross-basin approach: train the artificial neural network on data from a specific watershed and use another watershed as test data to predict. We found that this methodology was successful in predicting storage depletion in a particular watershed if the artificial neural network was trained on data from that same watershed. If the training database consisted of data of two watersheds combined, then the neural network was able to predict depletion in both watersheds. From here we concluded that the neural network was able to combine the characteristics of both watersheds and be used as a prediction tool for both, but not to other areas. Several concerns with the limitations of this methodology are discussed in chapter 2.

Chapter 3 tackled the extrapolation problem from a different angle by building a perceptual model that explains how water table depth is determined in different regions based on water

table depth data. Here we found that although in different regions/climates water table depth is predicted by the same predictor parameters (since all parameters were found to be relevant to some degree), the importance of each predictor varied per region or climate. The random forest predicting the global water table depth (trained on all the data) performed somewhat worse than the random forest models trained on data of subregions alone (for example Australia). Of course, it is significantly harder to encompass all possible climate regions or hydrological settings into one global model, but there is potential to weigh the importance of predictor parameters based on the location of the region that is predicted. For example, surface elevation, slope and drainage density were found to be most important for low relief areas and could be weighed heavier during calibration when being used in a numerical model.

Bibliography

Adnan, R. M., Liang, Z., Heddam, S., Zounemat-Kermani, M., Kisi, O., & Li, B. (2020). Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs. *Journal of Hydrology*, *586*, 124371.

Adnan, R. M., Liang, Z., Trajkovic, S., Zounemat-Kermani, M., Li, B., & Kisi, O. (2019). Daily streamflow prediction using optimally pruned extreme learning machine. *Journal of Hydrology*, *577*, 123981.

Adnan, R. M., Yuan, X., Kisi, O., & Yuan, Y. (2017). Streamflow forecasting using artificial neural network and support vector machine models. *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)*, *29*(1), 286–294.

Afzaal, H., Farooque, A. A., Abbas, F., Acharya, B., & Esau, T. (2020). Groundwater estimation from major physical hydrology components using artificial neural networks and deep learning. *Water*, *12*(1), 5.

Ahmad, S., & Simonovic, S. P. (2005). An artificial neural network model for generating hydrograph from hydro-meteorological parameters. *Journal of Hydrology*, *315*(1–4), 236–251.

Alcamo, J., Döll, P., Henrichs, T., Kaspar, F., Lehner, B., Rösch, T., & Siebert, S. (2003). Development and testing of the WaterGAP 2 global model of water use and availability. *Hydrological Sciences Journal*, *48*(3), 317–337.

Alpaydin, E. (2014). *Introduction to machine learning*. MIT press.

Al-Sudani, Z. A., Salih, S. Q., & Yaseen, Z. M. (2019). Development of multivariate adaptive regression spline integrated with differential evolution model for streamflow simulation. *Journal of Hydrology*, *573*, 1–12.

Añel, J. A., García-Rodríguez, M., & Rodeiro, J. (2019). Current status on the need for improved accessibility to climate change models. *Geoscientific Model Development Discussions*, 1–18.

Antonopoulos, V. Z., Gianniou, S. K., & Antonopoulos, A. V. (2016). Artificial neural networks and empirical equations to estimate daily evaporation: Application to lake Vegoritis, Greece. *Hydrological Sciences Journal*, *61*(14), 2590–2599.

Arnell, N. W. (1999). Climate change and global water resources. *Global Environmental Change*, *9*, S31–S49.

Barlow, P. M., & Leake, S. A. (2012). *Streamflow depletion by wells—Understanding and managing the effects of groundwater pumping on streamflow* (USGS Numbered Series No. 1376; Circular, p. 95). U.S. Geological Survey. http://pubs.er.usgs.gov/publication/cir1376

Bartholomé, E., & Belward, A. S. (2005). GLC2000: A new approach to global land cover mapping from Earth observation data. *International Journal of Remote Sensing*, *26*(9), 1959–1977. https://doi.org/10.1080/01431160412331291297

Batjes, N. H. (1997). A world dataset of derived soil properties by FAO–UNESCO soil unit for global modelling. *Soil Use and Management*, *13*(1), 9–16.

Beven, K. (2020). Deep learning, hydrological processes and the uniqueness of place. *Hydrological Processes*, *34*(16), 3608–3613.

Beven, K. J., & Chappell, N. A. (2021). Perceptual perplexity and parameter parsimony. *Wiley Interdisciplinary Reviews: Water*, *8*(4), e1530.

Bhasme, P., Vagadiya, J., & Bhatia, U. (2021a). Enhancing predictive skills in physically-consistent way: Physics Informed Machine Learning for Hydrological Processes. *ArXiv Preprint ArXiv:2104.11009*.

Bhasme, P., Vagadiya, J., & Bhatia, U. (2021b). Enhancing predictive skills in physically-consistent way: Physics Informed Machine Learning for Hydrological Processes. *ArXiv Preprint ArXiv:2104.11009*.

Blöschl, G. (2001). Scaling in hydrology. *Hydrological Processes*, *15*(4), 709–711.

Blöschl, G., Bierkens, M. F., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., Kirchner, J. W., McDonnell, J. J., Savenije, H. H., & Sivapalan, M. (2019). Twenty-three unsolved problems in hydrology (UPH)–a community perspective. *Hydrological Sciences Journal*, *64*(10), 1141–1158.

Blöschl, G., & Sivapalan, M. (1995). Scale issues in hydrological modelling: A review. *Hydrological Processes*, *9*(3–4), 251–290.

Bonaccorso, G. (2017). *Machine learning algorithms*. Packt Publishing Ltd.

Bouwer, H., & Maddock III, T. (1997). *Making sense of the interactions between groundwater and streamflow: Lessons for water masters and adjudicators*.

Bowden, G. J., Maier, H. R., & Dandy, G. C. (2005). Input determination for neural network models in water resources applications. Part 2. Case study: Forecasting salinity in a river. *Journal of Hydrology*, *301*(1–4), 93–107.

Bowes, B. D., Sadler, J. M., Morsy, M. M., Behl, M., & Goodall, J. L. (2019). Forecasting Groundwater Table in a Flood Prone Coastal City with Long Short-term Memory and Recurrent Neural Networks. *Water*, *11*(5), 1098. https://doi.org/10.3390/w11051098

Bredehoeft, J. D., Papadopoulos, S. S., & Cooper, H. H. (1982). The water budget myth (scientific basis of water management). *Studies in Geophysics, National Academy of Sciences*, 51–57.

Brown, S. C., Lester, R. E., Versace, V. L., Fawcett, J., & Laurenson, L. (2014). Hydrologic landscape regionalisation using deductive classification and random forests. *PloS One*, *9*(11), e112856.

Budyko, M. I. (1951). On climatic factors of runoff. *Prob. Fiz. Geogr*, *16*.

Cai, J., Liu, Y., Lei, T., & Pereira, L. S. (2007). Estimating reference evapotranspiration with the FAO Penman–Monteith equation using daily weather forecast messages. *Agricultural and Forest Meteorology*, *145*(1), 22–35. https://doi.org/10.1016/j.agrformet.2007.04.012

Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L., & Zdeborová, L. (2019). Machine learning and the physical sciences. *Reviews of Modern Physics*, *91*(4), 045002. https://doi.org/10.1103/RevModPhys.91.045002

Chang, F.-J., & Guo, S. (2020). Advances in Hydrologic Forecasts and Water Resources Management. *Water*, *12*(6), 1819. https://doi.org/10.3390/w12061819

Coulibaly, P., Anctil, F., & Bobée, B. (2000). Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. *Journal of Hydrology*, *230*(3–4), 244–257.

Crochemore, L., Isberg, K., Pimentel, R., Pineda, L., Hasan, A., & Arheimer, B. (2020). Lessons learnt from checking the quality of openly accessible river flow data worldwide. *Hydrological Sciences Journal*, *65*(5), 699–711.

Cudennec, C., Lins, H., Uhlenbrook, S., & Arheimer, B. (2020). *Editorial–Towards FAIR and SQUARE hydrological data*. Taylor & Francis.

Daliakopoulos, I. N., Coulibaly, P., & Tsanis, I. K. (2005). Groundwater level forecasting using artificial neural networks. *Journal of Hydrology*, *309*(1–4), 229–240.

Danielson, J. J., & Gesch, D. B. (n.d.). *Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010)*. 34.

de Graaf, I. E., Gleeson, T., van Beek, L. R., Sutanudjaja, E. H., & Bierkens, M. F. (2019). Environmental flow limits to global groundwater pumping. *Nature*, *574*(7776), 90–94.

De Graaf, I. E. M., Sutanudjaja, E. H., Van Beek, L. P. H., & Bierkens, M. F. P. (2015). A high-resolution global-scale groundwater model. *Hydrology and Earth System Sciences*, *19*(2), 823–837.

DeCoursey, D. G., & Deal, R. B. (1974). General aspects of multivariate analysis with applications to some problems in hydrology. *Proceedings of Symposium on Statistical Hydrology, USDA, Miscellaneous Publication*, *1275*, 47–68.

Dehghani, M., Saghafian, B., Nasiri Saleh, F., Farokhnia, A., & Noori, R. (2014). Uncertainty analysis of streamflow drought forecast using artificial neural networks and Monte-Carlo simulation. *International Journal of Climatology*, *34*(4), 1169–1180.

Deka, P. C. (2014). Support vector machine applications in the field of hydrology: A review. *Applied Soft Computing*, *19*, 372–386.

Di Gregorio, A. (2005). *Land cover classification system: Classification concepts and user manual: LCCS* (Vol. 2). Food & Agriculture Org.

Di Gregorio, A., & Jansen, L. J. (1998). A new concept for a land cover classification system. *The Land*, *2*(1), 55–65.

Dingman, S. L. (1978). Drainage density and streamflow: A closer look. *Water Resources Research*, *14*(6), 1183–1187.

Doherty, J. (2003). Ground water model calibration using pilot points and regularization. *Groundwater*, *41*(2), 170–177.

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

Domenico, P. A., & Schwartz, F. W. (1998). *Physical and chemical hydrogeology* (Vol. 506). Wiley New York.

Dyson, M., Bergkamp, G., & Scanlon, J. (2003). Flow: The essentials of environmental flows. *IUCN, Gland, Switzerland and Cambridge, UK*, 20–87.

EarthExplorer. (n.d.). Retrieved August 9, 2021, from https://earthexplorer.usgs.gov/

El Naqa, I., & Murphy, M. J. (2015). What is machine learning? In *Machine learning in radiation oncology* (pp. 3–11). Springer.

Fan, Y., Li, H., & Miguez-Macho, G. (2013). Global patterns of groundwater table depth. *Science*, *339*(6122), 940–943.

Feinstein, D. T., Fienen, M. N., Reeves, H. W., & Langevin, C. D. (2016). A Semi-Structured MODFLOW-USG Model to Evaluate Local Water Sources to Wells for Decision Support. *Groundwater*, *54*(4), 532–544.

Feinstein, D. T., Kauffman, L. J., Haserodt, M. J., Clark, B. R., & Juckem, P. F. (2018). *Extraction and development of inset models in support of groundwater age calculations for glacial aquifers*. US Geological Survey.

Fienen, M. N., Nolan, B. T., & Feinstein, D. T. (2016). Evaluating the sources of water to wells: Three techniques for metamodeling of a groundwater flow model. *Environmental Modelling & Software*, *77*, 95–107.

Fienen, M. N., Nolan, B. T., Feinstein, D. T., & Starn, J. J. (2015). *Metamodels to bridge the*

gap between modeling and decision support. http://digitalcommons.unl.edu/usgsstaffpub/860/

Fienen, M. N., & Plant, N. G. (2015). A cross-validation package driving Netica with python. *Environmental Modelling & Software*, *63*, 14–23.

Fraser, C. E., McIntyre, N., Jackson, B. M., & Wheater, H. S. (2013). Upscaling hydrological processes and land management change impacts using a metamodeling procedure. *Water Resources Research*, *49*(9), 5817–5833.

Gaume, E., & Gosset, R. (2003). Over-parameterisation, a major obstacle to the use of artificial neural networks in hydrology? *Hydrology and Earth System Sciences*, *7*(5), 693–706.

Gentine, P., Troy, T. J., Lintner, B. R., & Findell, K. L. (2012). Scaling in surface hydrology: Progress and challenges. *Journal of Contemporary Water Research & Education*, *147*(1), 28–40.

Gerten, D., Hoff, H., Rockström, J., Jägermeyr, J., Kummu, M., & Pastor, A. V. (2013). Towards a revised planetary boundary for consumptive freshwater use: Role of environmental flow requirements. *Current Opinion in Environmental Sustainability*, *5*(6), 551–558.

Gleeson, T., Alley, W. M., Allen, D. M., Sophocleous, M. A., Zhou, Y., Taniguchi, M., & VanderSteen, J. (2012). Towards sustainable groundwater use: Setting long-term goals, backcasting, and managing adaptively. *Groundwater*, *50*(1), 19–26.

Gleeson, T., Befus, K. M., Jasechko, S., Luijendijk, E., & Cardenas, M. B. (2016). The global volume and distribution of modern groundwater. *Nature Geoscience*, *9*(2), 161.

Gleeson, T., Moosdorf, N., Hartmann, J., & Van Beek, L. P. H. (2014). A glimpse beneath earth's surface: GLobal HYdrogeology MaPS (GLHYMPS) of permeability and porosity. *Geophysical Research Letters*, *41*(11), 3891–3898.

Gleeson, T., & Richter, B. (2018). How much groundwater can we pump and protect environmental flows through time? Presumptive standards for conjunctive management of aquifers and rivers. *River Research and Applications*, *34*(1), 83–92.

Gleeson, T., Wada, Y., Bierkens, M. F., & van Beek, L. P. (2012). Water balance of global aquifers revealed by groundwater footprint. *Nature*, *488*(7410), 197–200.

Gleeson, T., Wagener, T., Döll, P., Zipper, S. C., West, C., Wada, Y., Taylor, R., Scanlon, B., Rosolem, R., Rahman, S., Oshinlaja, N., Maxwell, R., Lo, M.-H., Kim, H., Hill, M., Hartmann, A., Fogg, G., Famiglietti, J. S., Ducharne, A., … Bierkens, M. F. P. (2021). GMD Perspective: The quest to improve the evaluation of groundwater representation in continental to global scale models. *Geoscientific Model Development Discussions*, 1–59. https://doi.org/10.5194/gmd-2021-97

Gleeson, T., Wang-Erlandsson, L., Zipper, S. C., Porkka, M., Jaramillo, F., Gerten, D., Fetzer, I., Cornell, S. E., Piemontese, L., & Gordon, L. J. (2020). The water planetary boundary: Interrogation and revision. *One Earth*, *2*(3), 223–234.

Glover, R. E., & Balmer, G. G. (1954). River depletion resulting from pumping a well near a river. *Eos, Transactions American Geophysical Union*, *35*(3), 468–470.

Gnann, S. J., Woods, R. A., & Howden, N. J. K. (2019). Is There a Baseflow Budyko Curve? *Water Resources Research*, *55*(4), 2838–2855. https://doi.org/10.1029/2018WR024464

Goh, G. B., Hodas, N. O., & Vishnu, A. (2017). Deep learning for computational chemistry. *Journal of Computational Chemistry*, *38*(16), 1291–1307.

Goldberg, D. E., & Holland, J. H. (1988). Genetic algorithms and machine learning. *Machine*

Learning, *3*(2), 95–99.
Govindaraju, R. S., & Rao, A. R. (2013). *Artificial neural networks in hydrology* (Vol. 36). Springer Science & Business Media.
Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science*, *24*, 395–419.
Grossberg, S. (1988). Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural Networks*, *1*(1), 17–61.
Gulli, A., & Pal, S. (2017). *Deep learning with Keras*. Packt Publishing Ltd.
Gupta, H. V. (2019). Transforming Earth Science by Bridging Machine Learning & Physics. *AGU Fall Meeting Abstracts*, *2019*, H33A-01.
Gupta, N., Rudra, R. P., & Parkin, G. (2006). Analysis of spatial variability of hydraulic conductivity at field scale. *Canadian Biosystems Engineering*, *48*, 1.
Haitjema, H., Kelson, V., & de Lange, W. (2001). Selecting MODFLOW cell sizes for accurate flow fields. *Groundwater*, *39*(6), 931–938.
Hall, C. A., Saia, S. M., Popp, A. L., Dogulu, N., Schymanski, S. J., Drost, N., van Emmerik, T., & Hut, R. (2021). A Hydrologist's Guide to Open Science. *Hydrology and Earth System Sciences Discussions*, 1–23.
Hantush, M. S. (1965). Wells near streams with semipervious beds. *Journal of Geophysical Research*, *70*(12), 2829–2838.
Harbaugh, A. W. (2005). *MODFLOW-2005, the US Geological Survey modular ground-water model: The ground-water flow process*. US Department of the Interior, US Geological Survey Reston. https://md.water.usgs.gov/gw/modflow/MODFLOW_Docs/TM6-A16-MODFLOW-2005.pdf
Harbaugh, A. W., Banta, E. R., Hill, M. C., & McDonald, M. G. (2000). Modflow-2000, the u. S. Geological survey modular ground-water model-user guide to modularization concepts and the ground-water flow process. *Open-File Report. U. S. Geological Survey*, *92*, 134.
Harris, I., Osborn, T. J., Jones, P., & Lister, D. (2020). Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Scientific Data*, *7*(1), 1–18.
Hengl, T., Mendes de Jesus, J., Heuvelink, G. B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., & Bauer-Marschallinger, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One*, *12*(2), e0169748.
Hernández, E., Sanchez-Anguix, V., Julian, V., Palanca, J., & Duque, N. (2016). Rainfall prediction: A deep learning approach. *International Conference on Hybrid Artificial Intelligence Systems*, 151–162.
Hill, M. C. (2000). Methods and guidelines for effective model calibration. In *Building Partnerships* (pp. 1–10).
Hill, M. C., & Tiedeman, C. R. (2006). *Effective groundwater model calibration: With analysis of data, sensitivities, predictions, and uncertainty*. John Wiley & Sons.
Hindman, M. (2015). Building Better Models: Prediction, Replication, and Machine Learning in the Social Sciences. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 48–62. https://doi.org/10.1177/0002716215570279
Hsieh, W. W. (2009). *Machine learning methods in the environmental sciences: Neural networks and kernels*. Cambridge university press.
Hunt, B. (1999). Unsteady stream depletion from ground water pumping. *Groundwater*, *37*(1), 98–102.

Huscroft, J., Gleeson, T., Hartmann, J., & Börker, J. (2018). Compiling and mapping global permeability of the unconsolidated and consolidated Earth: GLobal HYdrogeology MaPS 2.0 (GLHYMPS 2.0). *Geophysical Research Letters*, *45*(4), 1897–1904.

Iglesias, G., Kale, D. C., & Liu, Y. (2015). An examination of deep learning for extreme climate pattern analysis. *The 5th International Workshop on Climate Informatics*.

Kampichler, C., Wieland, R., Calmé, S., Weissenberger, H., & Arriaga-Weiss, S. (2010). Classification in conservation biology: A comparison of five machine-learning methods. *Ecological Informatics*, *5*(6), 441–450.

Kendy, E., & Bredehoeft, J. D. (2006). Transient effects of groundwater pumping and surface-water-irrigation returns on streamflow. *Water Resources Research*, *42*(8). http://onlinelibrary.wiley.com/doi/10.1029/2005WR004792/full

Khadri, S. F. R., & Pande, C. (2016). Ground water flow modeling for calibrating steady state using MODFLOW software: A case study of Mahesh River basin, India. *Modeling Earth Systems and Environment*, *2*(1), 39.

Khandelwal, A., Xu, S., Li, X., Jia, X., Stienbach, M., Duffy, C., Nieber, J., & Kumar, V. (2020). Physics Guided Machine Learning Methods for Hydrology. *ArXiv:2012.02854 [Physics]*. http://arxiv.org/abs/2012.02854

Konikow, L. F., & Leake, S. A. (2014). Depletion and capture: Revisiting "the source of water derived from wells." *Groundwater*, *52*(S1), 100–111.

Kraft, B., Jung, M., Körner, M., & Reichstein, M. (2020). Hybrid modeling: Fusion of a deep approach and physics-based model for global hydrological modeling. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *43*, 1537–1544.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12), 5089–5110.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.

Lange, H., & Sippel, S. (2020). Machine learning applications in hydrology. In *Forest-water interactions* (pp. 233–257). Springer.

Lee, S., Hyun, Y., Lee, S., & Lee, M.-J. (2020). Groundwater potential mapping using remote sensing and GIS-based machine learning techniques. *Remote Sensing*, *12*(7), 1200.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., & Aulagnier, S. (1996). Application of neural networks to modelling nonlinear relationships in ecology. *Ecological Modelling*, *90*(1), 39–52.

Lek, S., & Guégan, J.-F. (1999). Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling*, *120*(2–3), 65–73.

Li, H.-Y., Leung, L. R., Getirana, A., Huang, M., Wu, H., Xu, Y., Guo, J., & Voisin, N. (2015). Evaluating global streamflow simulations by a physically based routing model coupled with the community land model. *Journal of Hydrometeorology*, *16*(2), 948–971.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

Liu, Y., Racah, E., Correa, J., Khosrowshahi, A., Lavers, D., Kunkel, K., Wehner, M., & Collins, W. (2016). Application of deep convolutional neural networks for detecting extreme

weather in climate datasets. *ArXiv Preprint ArXiv:1605.01156*.
Llamas, M. R., & Custodio, E. (2002). *Intensive Use of Groundwater: Challenges and Opportunities*. CRC Press.
Mahesh, B. (2020). Machine Learning Algorithms-A Review. *International Journal of Science and Research (IJSR).[Internet]*, *9*, 381–386.
Maier, H. R., & Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. *Environmental Modelling & Software*, *15*(1), 101–124.
Mair, A., & Fares, A. (2010). Influence of groundwater pumping and rainfall spatio-temporal variation on streamflow. *Journal of Hydrology*, *393*(3–4), 287–308.
Makantasis, K., Karantzalos, K., Doulamis, A., & Doulamis, N. (2015). Deep supervised learning for hyperspectral data classification through convolutional neural networks. *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 4959–4962.
Malik, A., & Bhagwat, A. (2021). Modelling groundwater level fluctuations in urban areas using artificial neural network. *Groundwater for Sustainable Development*, *12*, 100484.
Marçais, J., & de Dreuzy, J.-R. (2017). Prospective Interest of Deep Learning for Hydrological Inference: J. Marçais and J.-R. de Dreuzy Groundwater xx, no. x: xx-xx. *Groundwater*, *55*(5), 688–692. https://doi.org/10.1111/gwat.12557
Maxwell, R. M., Condon, L. E., & Kollet, S. J. (2015). A high-resolution simulation of groundwater and surface water over most of the continental US with the integrated hydrologic model ParFlow v3. *Geoscientific Model Development*, *8*(3), 923–937.
Meyer, H., & Pebesma, E. (2020). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*.
Mitchell, T. M. 1951. (1997). *Machine Learning*. McGraw-Hill.
Mohan, C., Western, A. W., Wei, Y., & Saft, M. (2018). Predicting groundwater recharge for varying land cover and climate conditions–a global meta-study. *Hydrology and Earth System Sciences*, *22*(5), 2689–2703.
Montazeri, M., Montazeri, M., Montazeri, M., & Beigzadeh, A. (2016). Machine learning models in breast cancer survival prediction. *Technology and Health Care*, *24*(1), 31–42. https://doi.org/10.3233/THC-151071
Moradkhani, H., Hsu, K., Gupta, H. V., & Sorooshian, S. (2004). Improved streamflow forecasting using self-organizing radial basis function artificial neural networks. *Journal of Hydrology*, *295*(1–4), 246–262.
More, P. R. (2018). *Using Machine Learning to predict water table levels in a wet prairie in Northwest Ohio* [PhD Thesis]. Bowling Green State University.
Mudigonda, M., Kashinath, K., Racah, E., Mahesh, A., Liu, Y., Beckham, C., Biard, J., Kurth, T., Kim, S., & Kahou, S. (2021). Deep Learning for Detecting Extreme Weather Patterns. *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences*, 161–185.
Naghibi, S. A., Pourghasemi, H. R., & Dixon, B. (2016). GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environmental Monitoring and Assessment*, *188*(1), 1–27.
Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., & Gupta, H. V. (2021). What role does hydrological science play in the age of machine

learning? *Water Resources Research*, *57*(3), e2020WR028091.
Niswonger, R. G., Panday, S., & Ibaraki, M. (2011). *MODFLOW-NWT, a Newton formulation for MODFLOW-2005*. US Geological Survey. https://pubs.er.usgs.gov/publication/tm6A37
Niswonger, R. G., & Prudic, D. E. (2005). *Documentation of the Streamflow-Routing (SFR2) Package to include unsaturated flow beneath streams-A modification to SFR1*. US Geological Survey.
Noun Project: Free Icons & Stock Photos for Everything. (n.d.). Retrieved October 7, 2021, from https://thenounproject.com/
Parisouj, P., Mohebzadeh, H., & Lee, T. (2020). Employing machine learning algorithms for streamflow prediction: A case study of four river basins with different climatic zones in the United States. *Water Resources Management*, *34*(13), 4113–4131.
Pastore, M. (2018). Overlapping: A R package for estimating overlapping in empirical distributions. *Journal of Open Source Software*, *3*(32), 1023.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.
Peters-Lidard, C. D., Clark, M., Samaniego, L., Verhoest, N. E., Emmerik, T. van, Uijlenhoet, R., Achieng, K., Franz, T. E., & Woods, R. (2017). Scaling, similarity, and the fourth paradigm for hydrology. *Hydrology and Earth System Sciences*, *21*(7), 3701–3713.
Petty, T. R., & Dhingra, P. (2018). Streamflow hydrology estimate using machine learning (SHEM). *JAWRA Journal of the American Water Resources Association*, *54*(1), 55–68.
Poff, N. L., & Zimmerman, J. K. (2010). Ecological responses to altered flow regimes: A literature review to inform the science and management of environmental flows. *Freshwater Biology*, *55*(1), 194–205.
Pokhrel, Y. N., Koirala, S., Yeh, P. J.-F., Hanasaki, N., Longuevergne, L., Kanae, S., & Oki, T. (2015). Incorporation of groundwater pumping in a global L and S urface M odel with the representation of human impacts. *Water Resources Research*, *51*(1), 78–96.
Powers, S. M., & Hampton, S. E. (2019). Open science, reproducibility, and transparency in ecology. *Ecological Applications*, *29*(1), e01822.
Prabhat, M., Racah, E., Biard, J., Liu, Y., Mudigonda, M., Kashinath, K., Beckham, C., Maharaj, T., Kahou, S., & Pal, C. (2017). Deep Learning for Extreme Weather Detection. *AGU Fall Meeting Abstracts*, *2017*, IN11A-0022.
Prasad, P., Loveson, V. J., Kotha, M., & Yadav, R. (2020). Application of machine learning techniques in groundwater potential mapping along the west coast of India. *GIScience & Remote Sensing*, *57*(6), 735–752.
Rasouli, K., Hsieh, W. W., & Cannon, A. J. (2012). Daily streamflow forecasting by machine learning methods with weather and climate inputs. *Journal of Hydrology*, *414*, 284–293.
Razavi, S., & Gupta, H. V. (2016). A new framework for comprehensive, robust, and efficient global sensitivity analysis: 1. Theory. *Water Resources Research*, *52*(1), 423–439.
Razavi, S., Jakeman, A., Saltelli, A., Prieur, C., Iooss, B., Borgonovo, E., Plischke, E., Lo Piano, S., Iwanaga, T., Becker, W., Tarantola, S., Guillaume, J. H. A., Jakeman, J., Gupta, H., Melillo, N., Rabitti, G., Chabridon, V., Duan, Q., Sun, X., … Maier, H. R. (2021). The Future of Sensitivity Analysis: An essential discipline for systems modeling and policy support. *Environmental Modelling & Software*, *137*, 104954. https://doi.org/10.1016/j.envsoft.2020.104954

Razavi, S., & Tolson, B. A. (2013). An efficient framework for hydrologic model calibration on long data periods. *Water Resources Research*, *49*(12), 8418–8431.
Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *Science*, *331*(6018), 703–705.
Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., & Carvalhais, N. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204.
Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1
Reier Forradellas, R. F., Náñez Alonso, S. L., Jorge-Vazquez, J., & Rodriguez, M. L. (2021). Applied Machine Learning in Social Sciences: Neural Networks and Crime Prediction. *Social Sciences*, *10*(1), 4. https://doi.org/10.3390/socsci10010004
Reinecke, R., Foglia, L., Mehl, S., Herman, J. D., Wachholz, A., Trautmann, T., & Döll, P. (2019). Spatially distributed sensitivity of simulated global groundwater heads and flows to hydraulic conductivity, groundwater recharge, and surface water body parameterization. *Hydrology and Earth System Sciences*, *23*(11), 4561–4582.
Reinecke, R., Foglia, L., Mehl, S., Trautmann, T., Cáceres, D., & Döll, P. (2019). Challenges in developing a global gradient-based groundwater model (G 3 M v1. 0) for the integration into a global hydrological model. *Geoscientific Model Development*, *12*(6), 2401–2418.
Reinecke, R., Müller Schmied, H., Trautmann, T., Andersen, L. S., Burek, P., Flörke, M., Gosling, S. N., Grillakis, M., Hanasaki, N., & Koutroulis, A. (2021). Uncertainty of simulated groundwater recharge at different global warming levels: A global-scale multi-model ensemble study. *Hydrology and Earth System Sciences*, *25*(2), 787–810.
Reinecke, R., Wachholz, A., Mehl, S., Foglia, L., Niemann, C., & Döll, P. (2020). Importance of spatial resolution in global groundwater modeling. *Groundwater*, *58*(3), 363–376.
Ren, C., An, N., Wang, J., Li, L., Hu, B., & Shang, D. (2014). Optimal parameters selection for BP neural network based on particle swarm optimization: A case study of wind speed forecasting. *Knowledge-Based Systems*, *56*, 226–239.
Rockström, J., Steffen, W., Noone, K., Persson, \AAsa, Chapin III, F. S., Lambin, E., Lenton, T., Scheffer, M., Folke, C., & Schellnhuber, H. J. (2009). Planetary boundaries: Exploring the safe operating space for humanity. *Ecology and Society*, *14*(2).
Roshni, T., Jha, M. K., & Drisya, J. (2020). Neural network modeling for groundwater-level forecasting in coastal aquifers. *Neural Computing and Applications*, 1–18.
Sajedi-Hosseini, F., Malekian, A., Choubin, B., Rahmati, O., Cipullo, S., Coulon, F., & Pradhan, B. (2018). A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. *Science of the Total Environment*, *644*, 954–962.
Scheibe, T., & Yabusaki, S. (1998). Scaling of flow and transport behavior in heterogeneous groundwater systems. *Advances in Water Resources*, *22*(3), 223–238.
Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.
Schneider, A., Jost, A., Coulon, C., Silvestre, M., Théry, S., & Ducharne, A. (2017). Global-scale river network extraction based on high-resolution topography and constrained by lithology, climate, slope, and observed drainage density. *Geophysical Research Letters*, *44*(6), 2773–2781.
Shamshirband, S., Hashemi, S., Salimi, H., Samadianfard, S., Asadi, E., Shadkani, S., Kargar, K.,

Mosavi, A., Nabipour, N., & Chau, K.-W. (2020). Predicting standardized streamflow index for hydrological drought using machine learning models. *Engineering Applications of Computational Fluid Mechanics*, *14*(1), 339–350.

Shanmuganathan, S. (2016). Artificial neural network modelling: An introduction. In *Artificial neural network modelling* (pp. 1–14). Springer.

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, *54*(11), 8558–8593.

Shen, C., Chen, X., & Laloy, E. (2021). Broadening the Use of Machine Learning in Hydrology. *Frontiers in Water*, *3*, 38.

Shen, C., & Lawson, K. (2021). Applications of Deep Learning in Hydrology. *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences*, 283–297.

Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D.-Y., Wong, W., & Woo, W. (2017). Deep learning for precipitation nowcasting: A benchmark and a new model. *ArXiv Preprint ArXiv:1706.03458*.

Slater, L. J., Thirel, G., Harrigan, S., Delaigue, O., Hurley, A., Khouakhi, A., Prosdocimi, I., Vitolo, C., & Smith, K. (2019). Using R in hydrology: A review of recent developments and future directions. *Hydrology and Earth System Sciences*, *23*(7), 2939–2963.

Snyder, W. M. (1962). Some possibilities for multivariate analysis in hydrologic studies. *Journal of Geophysical Research*, *67*(2), 721–729.

Sobieraj, J. A., Elsenbeer, H., & Cameron, G. (2004). Scale dependency in spatial patterns of saturated hydraulic conductivity. *Catena*, *55*(1), 49–77.

Sperotto, A., Molina, J. L., Torresan, S., Critto, A., Pulido-Velazquez, M., & Marcomini, A. (2019). A Bayesian Networks approach for the assessment of climate change impacts on nutrients loading. *Environmental Science & Policy*, *100*, 21–36.

Sreekanth, P. D., Geethanjali, N., Sreedevi, P. D., Ahmed, S., Kumar, N. R., & Jayanthi, P. K. (2009). Forecasting groundwater level using artificial neural networks. *Current Science*, 933–939.

Stagge, J. H., Rosenberg, D. E., Abdallah, A. M., Akbar, H., Attallah, N. A., & James, R. (2019). Assessing data availability and research reproducibility in hydrology and water resources. *Scientific Data*, *6*(1), 1–12.

Stathakis, D. (2009). How many hidden layers and nodes? *International Journal of Remote Sensing*, *30*(8), 2133–2147.

Sternberg, M. J. E., King, R. D., Lewis, R. A., Muggleton, S., Bodmer, W. F., & Donnelly, P. J. (1994). Application of machine learning to structural molecular biology. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *344*(1310), 365–371. https://doi.org/10.1098/rstb.1994.0075

Sutanudjaja, E. H., van Beek, L. P. H., de Jong, S. M., van Geer, F. C., & Bierkens, M. F. P. (2011). Large-scale groundwater modeling using global datasets: A test case for the Rhine-Meuse basin. *Hydrology and Earth System Sciences*, *15*(9), 2913–2935. https://doi.org/10.5194/hess-15-2913-2011

Sutanudjaja, E. H., Van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H., Drost, N., Van Der Ent, R. J., De Graaf, I. E., Hoch, J. M., & De Jong, K. (2018). PCR-GLOBWB 2: A 5 arcmin global hydrological and water resources model. *Geoscientific Model Development*, *11*(6), 2429–2453.

Taormina, R., & Chau, K. (2015). Neural network river forecasting with multi-objective fully informed particle swarm optimization. *Journal of Hydroinformatics*, *17*(1), 99–113.

Tarboton, D. G., Bras, R. L., & Rodriguez-Iturbe, I. (1992). A physical basis for drainage density. *Geomorphology*, *5*(1–2), 59–76.

Thessen, A. (2016). Adoption of machine learning techniques in ecology and earth science. *One Ecosystem*, *1*, e8621.

Tokar, A. S., & Johnson, P. A. (1999). Rainfall-runoff modeling using artificial neural networks. *Journal of Hydrologic Engineering*, *4*(3), 232–239.

Tripathi, S., Srinivas, V. V., & Nanjundiah, R. S. (2006). Downscaling of precipitation for climate change scenarios: A support vector machine approach. *Journal of Hydrology*, *330*(3–4), 621–640.

Trippi, R. R., & Turban, E. (1992). *Neural networks in finance and investing: Using artificial intelligence to improve real world performance*. McGraw-Hill, Inc.

Wada, Y., & Bierkens, M. F. (2014). Sustainability of global water use: Past reconstruction and future projections. *Environmental Research Letters*, *9*(10), 104003.

Wada, Y., Van Beek, L. P., Van Kempen, C. M., Reckman, J. W., Vasak, S., & Bierkens, M. F. (2010). Global depletion of groundwater resources. *Geophysical Research Letters*, *37*(20).

Wagener, T. (2020). On doing Hydrology with Lions. *EGU General Assembly Conference Abstracts*, 9924.

Wagener, T., Gleeson, T., Coxon, G., Hartmann, A., Howden, N., Pianosi, F., Rahman, M., Rosolem, R., Stein, L., & Woods, R. (2021). On doing hydrology with dragons: Realizing the value of perceptual models and knowledge accumulation. *WIREs Water*, *n/a*(n/a), e1550. https://doi.org/10.1002/wat2.1550

Wanas, N., Auda, G., Kamel, M. S., & Karray, F. (1998). On the optimal number of hidden nodes in a neural network. *Conference Proceedings. IEEE Canadian Conference on Electrical and Computer Engineering (Cat. No. 98TH8341)*, *2*, 918–921.

Willard, J., Jia, X., Xu, S., Steinbach, M., & Kumar, V. (2020). Integrating physics-based modeling with machine learning: A survey. *ArXiv Preprint ArXiv:2003.04919*.

Winter, T. C. (2001). The concept of hydrologic landscapes. *JAWRA Journal of the American Water Resources Association*, *37*(2), 335–349.

Xu, T., & Liang, F. (2021). Machine learning for hydrologic sciences: An introductory overview. *Wiley Interdisciplinary Reviews: Water*, e1533.

Yang, T., Sun, F., Gentine, P., Liu, W., Wang, H., Yin, J., Du, M., & Liu, C. (2019). Evaluation and machine learning improvement of global hydrological model-based flood simulations. *Environmental Research Letters*, *14*(11), 114027.

Zealand, C. M., Burn, D. H., & Simonovic, S. P. (1999). Short term streamflow forecasting using artificial neural networks. *Journal of Hydrology*, *214*(1–4), 32–48.

Zhang, L., Zhang, L., & Du, B. (2016). Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, *4*(2), 22–40.

Zhang, P., Zhang, L., Leung, H., & Wang, J. (2017). A deep-learning based precipitation forecasting approach using multiple environmental factors. *2017 IEEE International Congress on Big Data (BigData Congress)*, 193–200.

Zhao, T., Zhu, Y., Ye, M., Mao, W., Zhang, X., Yang, J., & Wu, J. (2020). Machine-Learning Methods for Water Table Depth Prediction in Seasonal Freezing-Thawing Areas. *Groundwater*, *58*(3), 419–431.

Zhou, Y., & Li, W. (2011). A review of regional groundwater flow modeling. *Geoscience Frontiers*, *2*(2), 205–214.

Zipper, S. C., Dallemagne, T., Gleeson, T., Boerman, T. C., & Hartmann, A. (2018). Groundwater pumping impacts on real stream networks: Testing the performance of simple management tools. *Water Resources Research*, *54*(8), 5471–5486.

Zipper, S. C., Gleeson, T., Kerr, B., Howard, J. K., Rohde, M. M., Carah, J., & Zimmerman, J. (2019). Rapid and accurate estimates of streamflow depletion caused by groundwater pumping using analytical depletion functions. *Water Resources Research*, *55*(7), 5807–5829.

Zipper, S. C., Stack Whitney, K., Deines, J. M., Befus, K. M., Bhatia, U., Albers, S. J., Beecher, J., Brelsford, C., Garcia, M., & Gleeson, T. (2019). Balancing open science and data privacy in the water sciences. *Water Resources Research*, *55*(7), 5202–5211.

Zlotnik, V. A. (2004). A concept of maximum stream depletion rate for leaky aquifers in alluvial valleys. *Water Resources Research*, *40*(6).

Zlotnik, V. A., Huang, H., & Butler Jr, J. J. (1999). *Evaluation of stream depletion considering finite stream width, shallow penetration, and properties of streambed sediments*.

Zlotnik, V. A., & Tartakovsky, D. M. (2008). Stream depletion by groundwater pumping in leaky aquifers. *Journal of Hydrologic Engineering*, *13*(2), 43–50.

Appendices

Appendix A - Supplementary information Chapter 2: "Predicting the transient sources of groundwater abstraction using artificial neural networks: possibilities and limitations"

A-1 Finding optimal local area size

The methodology of using local areas as independent samples allow for the constructing of an extensive database if the local areas are found to be independent. Equation A-1 shows the mass balance used in the current study.

$$SHALGW + DEEPGW + SURFW = Q_{well} \qquad (A-1)$$

(1) **Shallow storage** (*SHALGW*): groundwater from storage release in the shallowest part of the system (<100 ft below ground surface). In the numerical model this is calculated as the change in storage volume of all the cells in model layer 1 within the local area, that is, to a depth no greater than 100 ft below land surface.
(2) **Deep storage** (*DEEPGW*): groundwater from storage flowing into layer 1 from deeper model layers, calculated in the model as the change in volumetric flux from layer 2 to layer 1.
(3) **Surface water** (*SURFW):* change in volumetric flux between surface water features and the aquifer, representing water captured from surface water sources. SURFW is the sum of water coming from streams (streamflow routing cells), lakes and wetlands (drain cells)

It must be stated that this mass balance threshold is met for certain specific circumstances. This threshold is met for a combination of: (1) well spacing, (2) local area size, (3) pumping rate, and (4) pumping time. When pumping rate and pumping time are chosen differently the mass balance threshold may not be reached. As is stated in the main text, horizontal groundwater flow is expected to be a significant source of water when either pumping time or pumping rate is increased. We reason that under our current conditions (combination of well spacing, local area size, pumping rate and pumping time) the cone of depression has moved far enough from the pumping well to cause significant drawdown near the edges of the local area, causing horizontal flow of groundwater into the local area.

For the current study we found that for the KALA, UPFOX and MANI databases, we had to remove up to 40% of local area samples out of the database for not meeting the mass balance threshold. This was especially the case for the MANI database (~40%) and less for the KALA and UPFOX database (~15 and 20% respectively).

Each model for each basin was run for a range of pumping rates (100, 1000, 10,000, 100,000, 1,000,000 ft^3/day) and well spacings (25, 50, 75 and 100 cells) on with stress periods lengths

of 1 month, 6 months, 1 year, 10 years, 25 years, 50 years and 100 years. The general trend over the different models, with exception of the UPFOX model, is that the mass balance holds up very well (> 80% of samples) for instances where (1) the well spacing is 50 cells or higher, (2) local size 25 cells or higher, (2) pumping rates are between 10,000 and 100,000 ft^3/day, and/or (3) pumping times are lower than less than 50 years. Having a simulation time longer than 25 years or a pumping rate higher than 100,000 ft^3/day, causes the mass balance threshold to be failed to meet. When a lower pumping rate was used (<10,000 ft^3/day) we found that the amount of the changes in baseflow became negligible. That is also the reason why we chose to use a combined response variable of surface water (SURFW) instead of treating wetlands and streamflow as different responses.

The local area size used in this study was also used in previous work by Feinstein et al. (Feinstein et al., 2016) and Fienen et al. (Fienen et al., 2016). The models in this study are inset models of the larger Michigan basin model of the USGS and have the same cell size as in the previous studies.

A-2 Neural network theory and optimization

A-2.1 Neural network theory

A neural network consists of nodes fitted into distinct layers and are mathematically connected (often linearly). A neural network consists of an *input layer* with one node per input parameter, an *output layer* with one node per response and several layers of nodes in between, also known as *hidden layers*. The number of hidden layers and the number of nodes per hidden layer is variable and dependent on the database and the scope of the problem. Various approaches have been examined for optimizing the number of hidden nodes per layer (Stathakis, 2009; Wanas et al., 1998).

The difference between the calculated activation and the actual value is called the *error* or *cost*. The total cost function *J* of the neural network is calculated by the log-cosh function (Shanmuganathan & Samarasinghe, 2016):

$$J = \sum_{i=0}^{n} log\ log\ \left(cosh\ cosh\left(y_{pred}^{(i)} - y^{(i)}\right)\right) \qquad (A-2)$$

Here $y_{pred}^{(i)}$ is the value for y predicted by the neural network in output node i, $y^{(i)}$,is the corresponding real value for the same output node i. The total cost J is then calculated over the total number of nodes in the output layer. Neural network training is the process of minimizing the cost function through the process of backpropagation (Lek & Guégan, 1999; Schmidhuber, 2015), which updates all weights and biases and iteratively recalculates the cost function until it has found a minimum value. This minimum cost configuration represents the configuration of weights within the weight matrix which gives the best mathematical relation between the input and output data.

A-2.2 Activation function

All nodes in a neural network represent a value typically in the range of $[0,1]$ (also called its *activation*). The activation of a single node in the next layer is calculated by the sum of all the activations in the previous layer, multiplied by a *weight* parameter and a *bias* term. This is mathematically summarized in equation A-3:

$$a_n^{(l)} = g\left(\beta + \sum_{i=0}^{K} \theta_i a_i^{(l-1)}\right) \qquad (A-3)$$

Where $a_i^{(l)}$ is the activation of the *i'th* node in layer *l*, $z_i^{(l-1)}$ is the activation of the *i'th* node in the previous layer (which has a number of nodes equal to *K* (excluding the bias term)), θ_i is the weight factor associated with $a_i^{(l-1)}$ and β is called the *bias term* and is set to be equal to 1. Weight values are initially randomly chosen but are later updated by the model in the process of backpropagation. The activation is then calculated by substituting the sum into an activation function. The additional benefit of this procedure is that this allows for input features of different dimensions or orders of magnitude. Both the Rectified Linear Unit (ReLU) function (equation A-1) and the Sigmoid function (equation A-2) are used in this study as activation functions (Schmidhuber, 2015). For the ReLU function the activation equals z for any z > 0 and equals 0 for z ≤ 0. The Sigmoid function scales all values between [0,1].

$$g(z) = (0, z) \qquad (A-4)$$

$$g(z) = \frac{1}{1 + e^{-z}} \qquad (A-5)$$

The activations are calculated forward into the model until it reaches the output layer of the model (through the process of *forward propagation*). The calculated activation of an output node is compared to the actual value.

A-2.3 Cost function

In this study three common cost functions used in regression problems (Mean Squared Error [MSE], Mean Absolute Error [MAE] and Logarithmic Hyperbolic Cosine [Log-Cosh]) (Shanmuganathan, 2016) were tested. The Log-Cosh activation function shows the lowest overall cost for all datasets.

$$MSE = \frac{\sum_{i=1}^{n} (y_{pred}^{(i)} - y_{true}^{(i)})^2}{n} \qquad (A-6)$$

$$MAE = \frac{\sum_{i=1}^{n} \left| y_{pred}^{(i)} - y_{true}^{(i)} \right|}{n} \qquad (A-7)$$

$$LOGCOSH = \sum_{i=1}^{n} log\left(cosh\, cosh\left(y_{pred}^{(i)} - y_{true}^{(i)}\right)\right) \qquad (A-8)$$

Here $y_{pred}^{(i)}$ is the value for y predicted by the neural network in output node i, $y_{true}^{(i)}$, is the corresponding real value for the same output node i. The total cost is then calculated over the total number of nodes in the output layer. Neural network training is the process of minimizing the cost function through the process of backpropagation (Lek & Guégan, 1999; Schmidhuber, 2015), which updates all weights and biases and iteratively recalculates the cost function until it has found a minimum value. This minimum cost configuration represents the configuration of weights within the weight matrix which gives the best mathematical relation between the input and output data.

A-3 Database and neural network

Table A-1: Representative HUC8 basins and size of each individual database

Name	HUC8 basins	Number of samples
KALA	KALA	2878
MANI	MANI	875
UPFOX	UPFOX	1703
KAMA	KALA and MANI	3753
MAUP	MANI and UPFOX	2578
KAUP	KALA and UPFOX	4581
KAMAUP	KALA, MANI and UPFOX	5456

Table A-2: Optimized neural network structure

Parameter	
Number of input nodes	7 (for base model)
Number of hidden layers	6
Number of nodes in hidden layers 1 through 6	100
Number of output nodes	5
Activation function	ReLU
Cost function	LogCosh
Optimizer (keras)	Adam
Learning rate	0.0001 - 0.005
Number of epochs	500 - 1000

Appendix B - Supplementary information Chapter 3: "Disentangling process controls on global groundwater table depth patterns using random forests"

Random Forest training was performed in Python using the `Scikit-learn` package (Pedregosa et al., 2011). Apart from the selection of suitable predictor parameters the most important tuning parameters for a random forest model include the number of trees (`n_estimators`), the maximum number of levels in the tree (`max_depth`), the number of features to consider at each split (`max_features`), the minimum number of data samples required to split a node (`min_samples_split`), and the minimum samples required on each leaf node (`min_samples_leaf`). Discrete parameters were split using one-hot encoding.

Table B-1 Fan et al groundwater database data (original and resampled to 0.1° spatial resolution)

Dataset (per continent)	**# of data points (original)**	**# of data points (resampled to 0.1°)**
North America	1,375,059	55428
South America	34,174	8536
Europe	77,287	4330
Asia	1,099	621
Africa	430	236
Oceania (incl. Australia)		10775
Total	1,584,393	

Table B-2 Database split based on geographic regions (Figure 3.4) and the percentage of area coverage by both the training data (domain) and the unknown (predicted) area domain

Database split	**Size (# of samples)**	**% coverage training domain**	**% coverage predicted area**
Rhine Delta	954	100	0
France	1686	26.34	73.66
Iberia	766	12.64	87.36
USA	39863	48.72	51.28
California	1276	31.06	68.94
Colorado	1456	51.72	48.28
Brazil	8320	11.83	88.17

Australia	10774	15.68	84.32
Global	79880	6.96	93.04

Table B-3 Results of overlap index analysis

Predictor	Symbol	Data type	η (0.5°)	η (0.25°)	η (0.1°)	η (0.05°)	Mean (standard devation)
Surface elevation	***SurfElev***	Continuous	0.83	0.8201	0.7842	0.7567	**0.798 (0.029)**
Log(Permeability)	***Perm***	Discrete	0.6537	0.5955	0.5188	0.4564	**0.556 (0.075)**
porosity	***Por***	Discrete	0.6726	0.6049	0.5182	0.4556	**0.563 (0.083)**
drainage density	***DrainDens***	Continuous	0.8608	0.8113	0.7587	0.7091	**0.785 (0.057)**
Monthly precipitation	***Precip***	Continuous	0.7263	0.7254	0.7118	0.6759	**0.71 (0.02)**
Potential evapotranspiration	***PET***	Continuous	0.6005	0.5493	0.5068	0.4587	**0.529 (0.052)**
Diurnal temperature range	***ΔT***	Continuous	0.8373	0.7965	0.7422	0.6928	**0.767 (0.055)**
Land use type	***LandUse***	Discrete	0.7246	0.6666	0.6145	0.5353	**0.635 (0.07)**
Dominant soil group	***DomSoil***	Discrete	0.7263	0.6587	0.5655	0.4817	**0.608 (0.09)**
		Mean (std)	**0.737 (0.08)**	**0.692 (0.1)**	**0.636 (0.11)**	**0.58 (0.12)**	

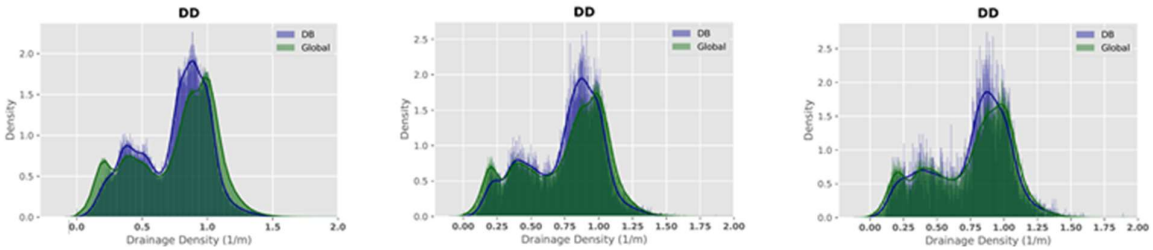

Figure B-1 Example of kernel distributions for predictor parameter DD within the Fan et al. database (database, blue) and the global distribution (Global). The overlap index (OI) is calculated as the percentage overlap between the areas under the kernel density distribution curves. The three graphs (from left to right) show the results for 0.1°, 0.25° and 0.5° spatial resolution.

Predictor parameter independence

Machine learning algorithm performance suffers when one or more chosen predictor parameters are either unrelated to the target parameter or if one or more either give redundant or contradicting information. To check whether the chosen predictor parameters are independent of each other we have plotted all continuous predictor parameters (pairwise) and calculated the Pearson r correlation coefficients.

Table B-4 Pearson R coefficients for all pairwise predictor comparisons

Pearson r	**SurfElev**	**Slope**	**DrainDens**	**Precip**	**PET**	**Delta T**
SurfElev	-	-	-	-	-	-
Slope	0.51	-	-	-	-	-
DrainDens	0.16	0.31	-	-	-	-
Precip	-0.18	0.13	0.3	-	-	-
PET	-0.14	-0.04	-0.31	0.02	-	-
Delta T	0.31	0.01	-0.22	-0.35	0.49	-

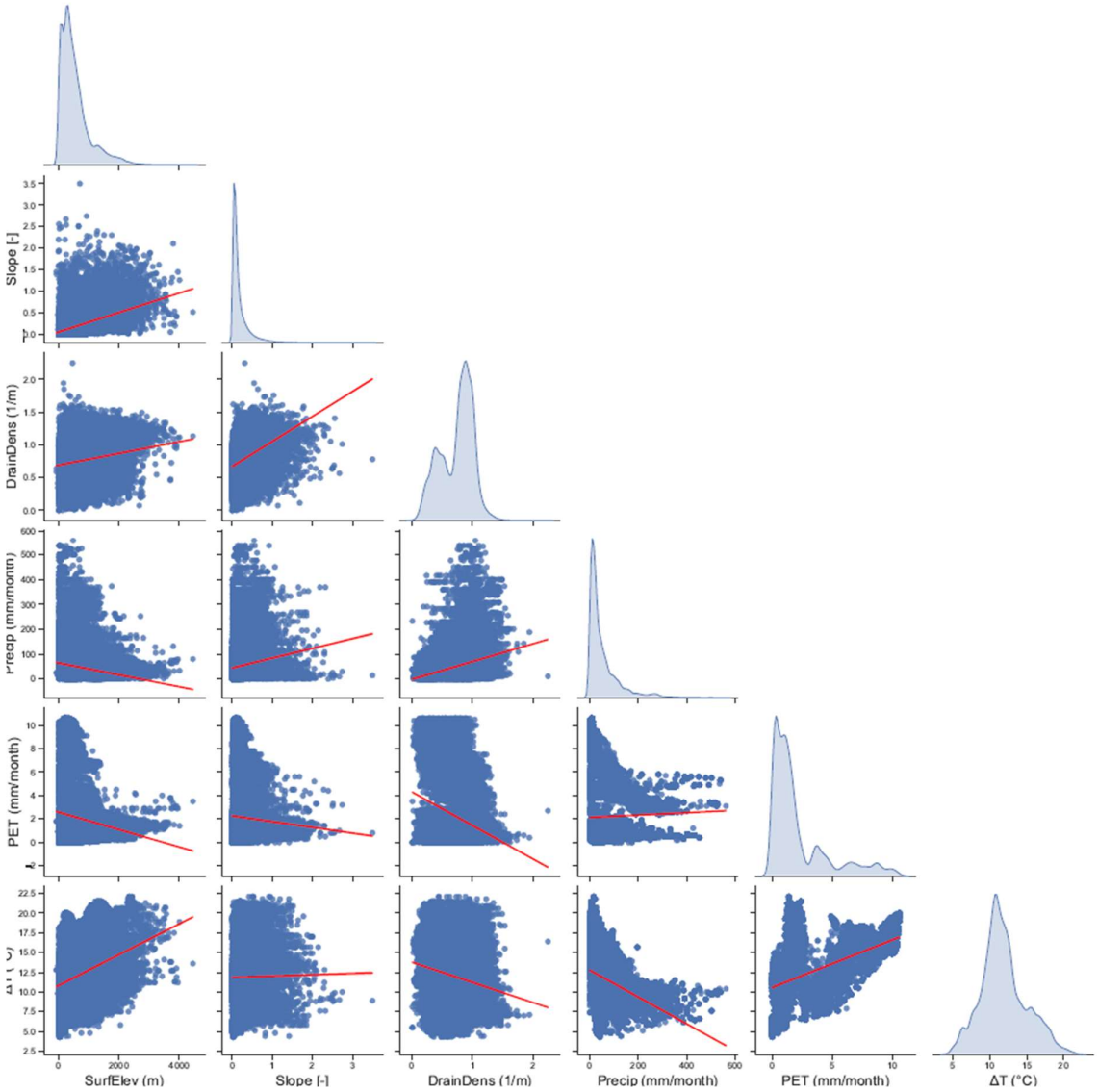

Figure B-2 Predictor parameter correlations and best fit trend lines for continuous predictor variables. Diagonal graphs show kernel density distributions of these parameters. Corresponding pearson R coefficients are given in table B-4.

Based on the results in table B-2 that η is generally higher for continuous data compared to discrete data. This can be explained by the fact that histograms for continuous data are approximate the kernel distribution function for large numbers of bins. For discrete data the number of bins is predetermined, and the kernel distribution function is generally more different from the histogram for discrete data.

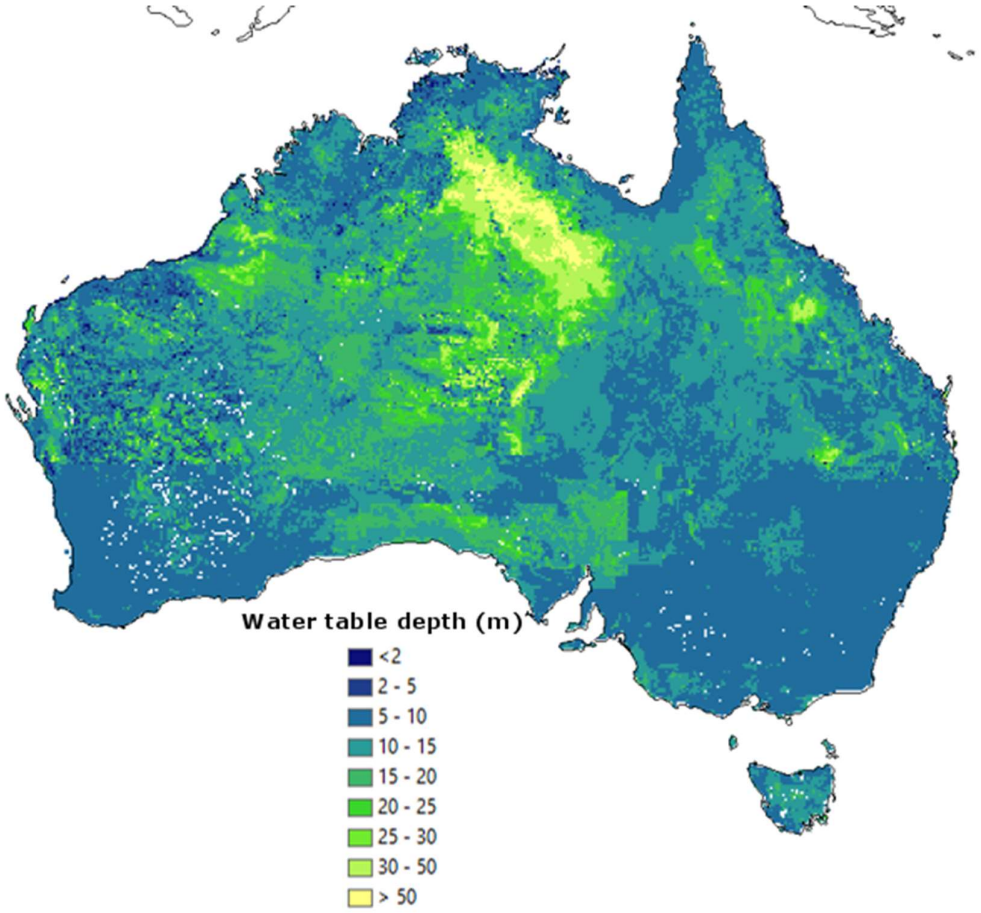

Figure B-3 Combined water table depth map of Australia using Fan et al data and predicted water table depth by random forests

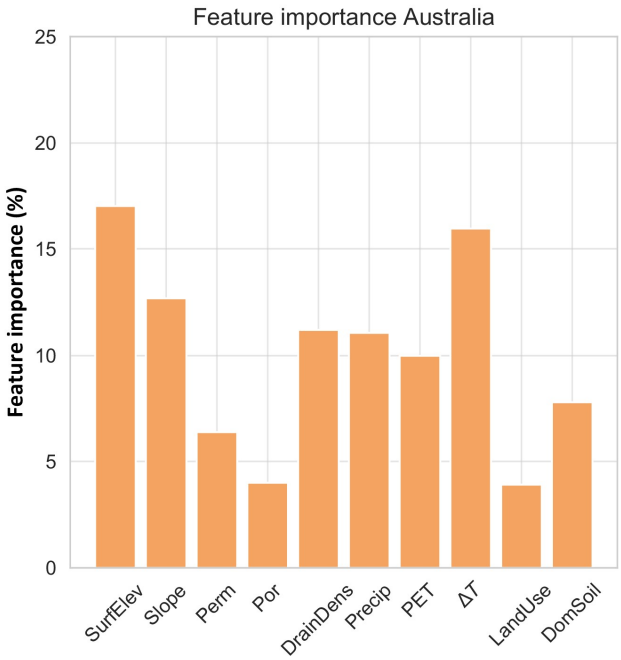

Figure B-4 Feature importance from the Australia random forest model

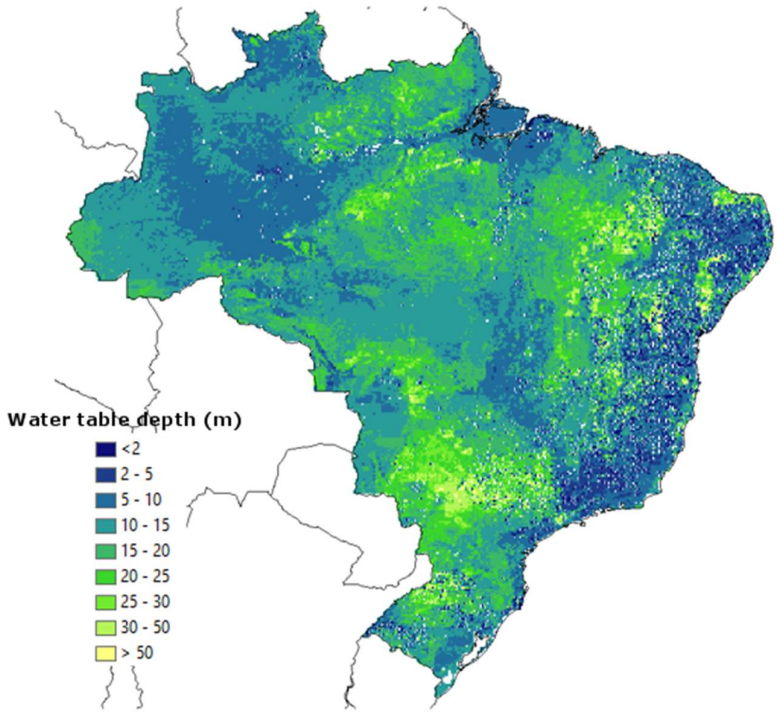

Figure B-5 Combined water table depth map of Brazil using Fan et al data and predicted water table depth by random forests

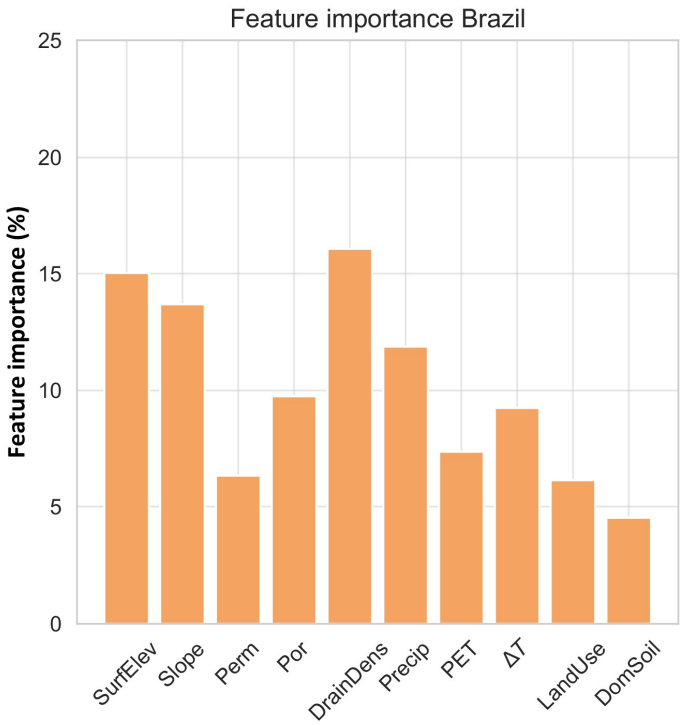

Figure B-6 Feature importance from the Brazil random forest model

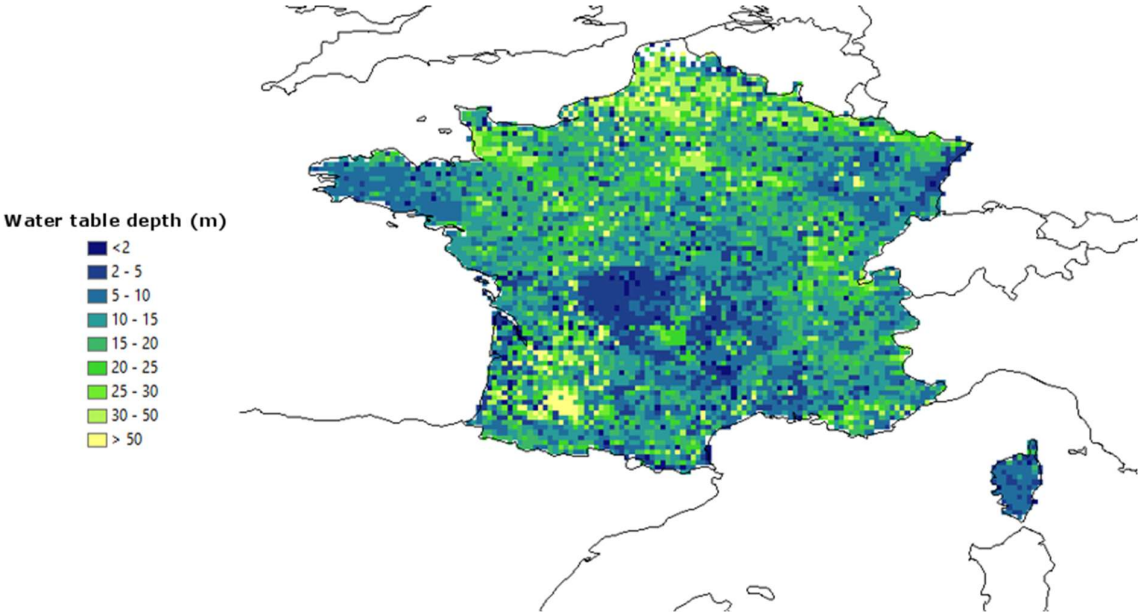

Figure B-7 Combined water table depth map of France using Fan et al data and predicted water table depth by random forests

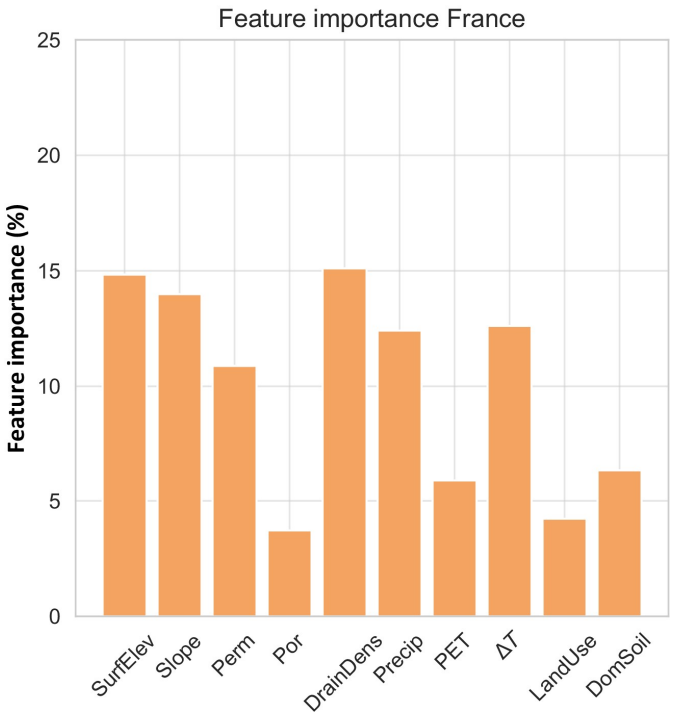

Figure B-8 Feature importance from the France random forest model

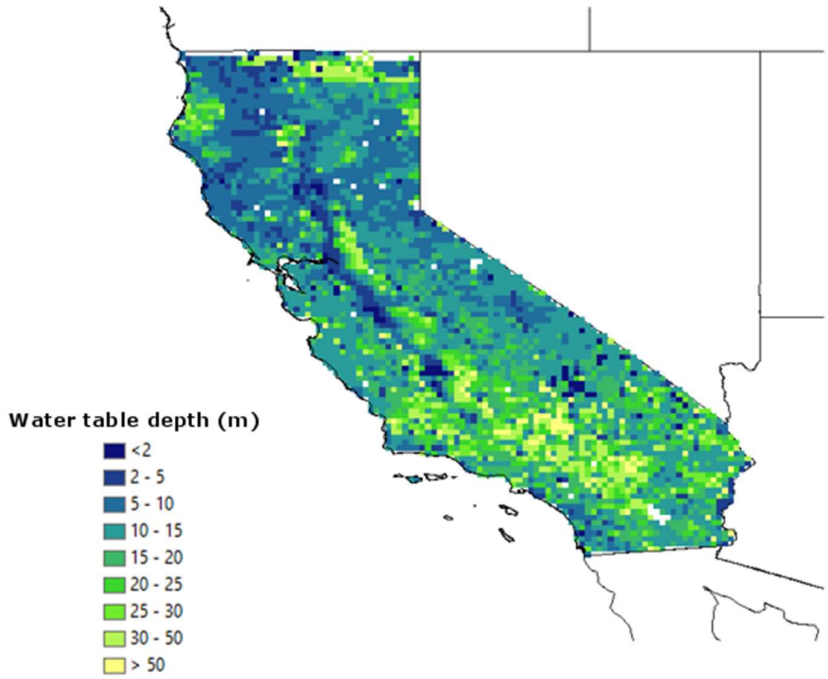

Figure B-9 Combined water table depth map of California using Fan et al data and predicted water table depth by random forests

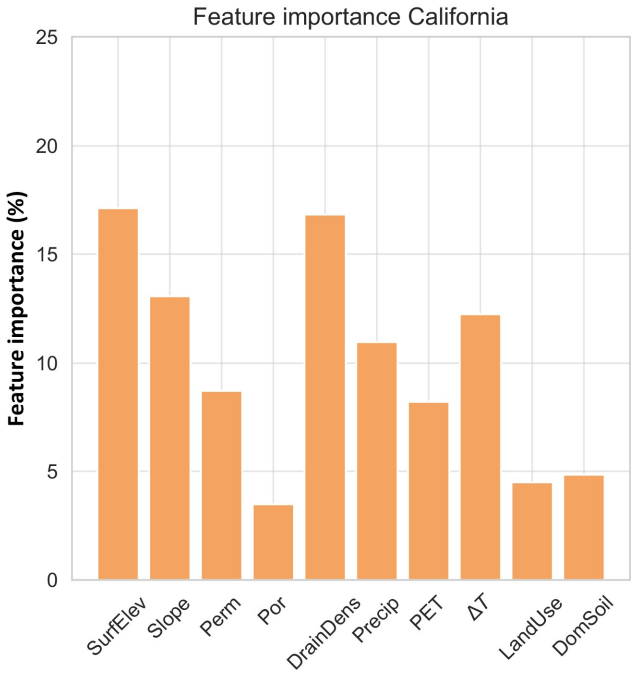

Figure B-10 Feature importance from the California random forest model

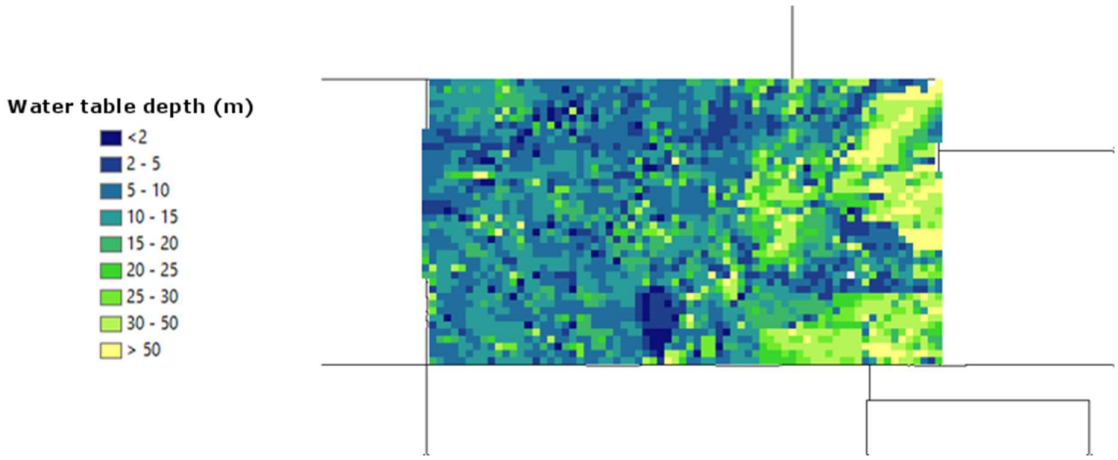

Figure B-11 Combined water table depth map of Colorado using Fan et al data and predicted water table depth by random forests

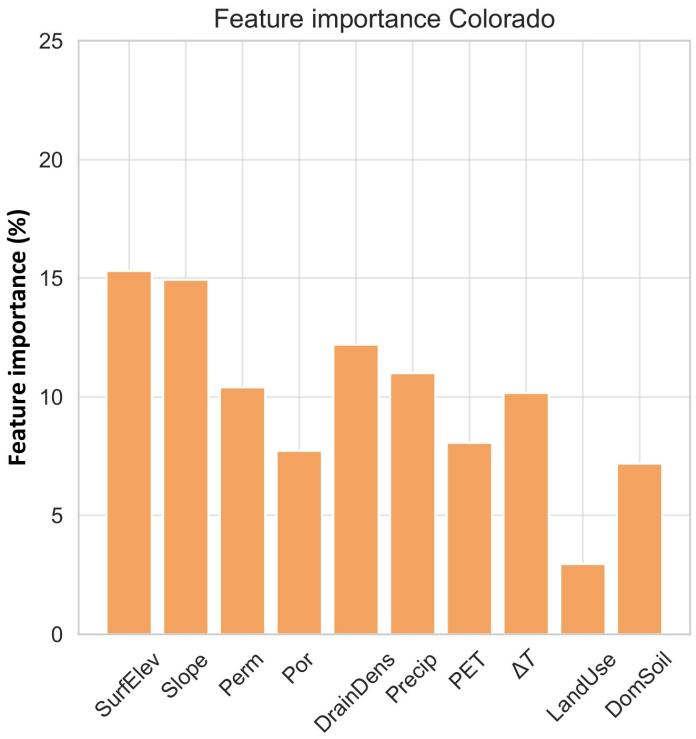

Figure B-12 Feature importance from the Colorado random forest model

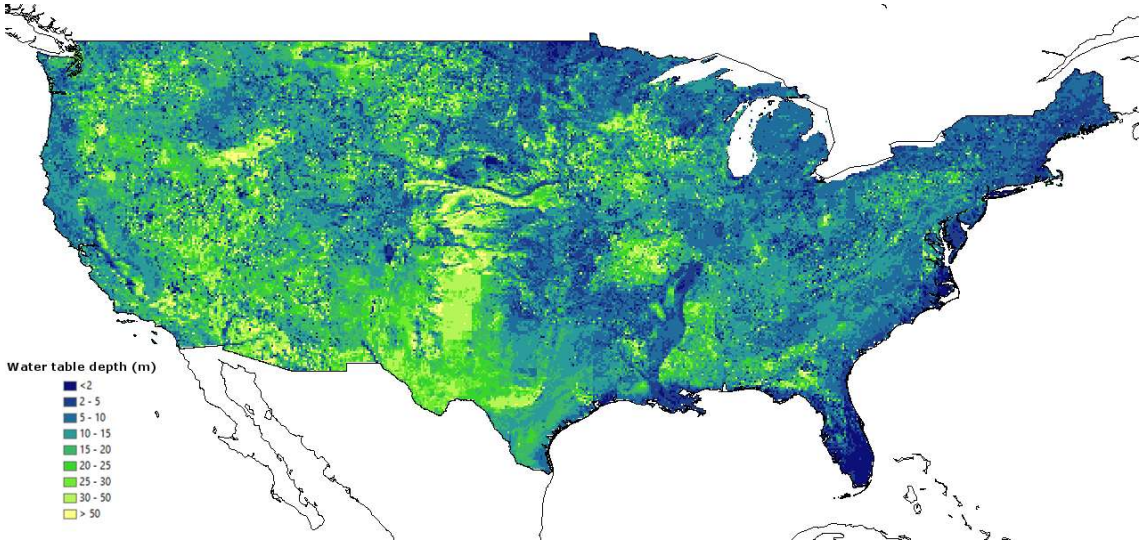

Figure B-13 Combined water table depth map of Australia using Fan et al data and predicted water table depth by random forests

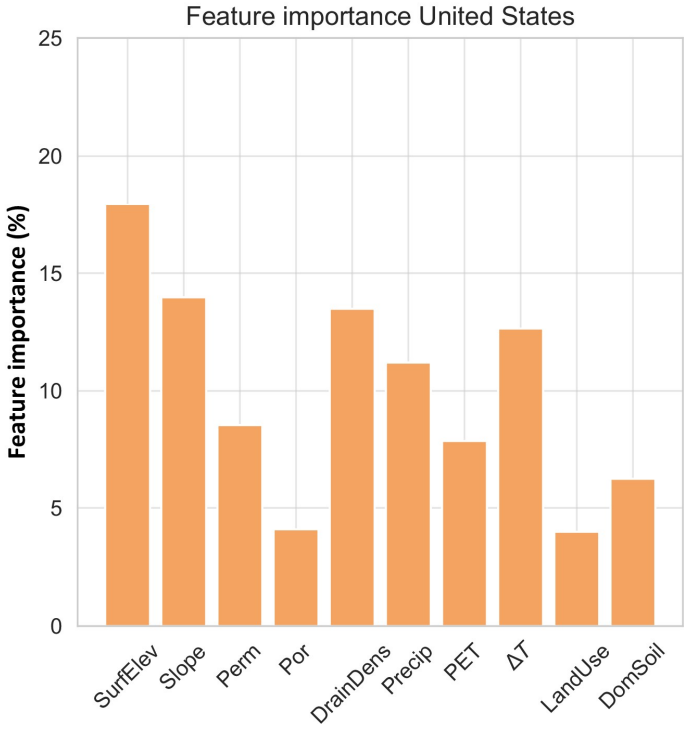

Figure B-14 Feature importance from the USA random forest model

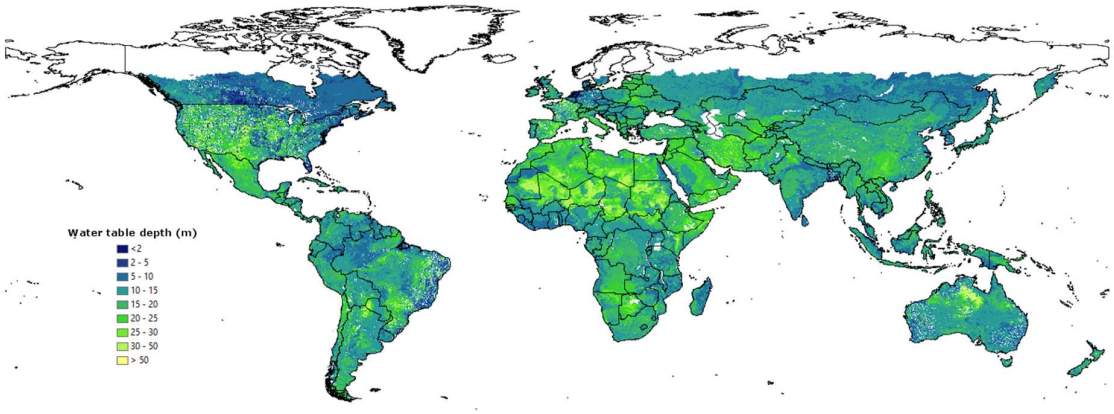

Figure B-15 Combined water table depth map of the Global model using Fan et al data and predicted water table depth by random forests

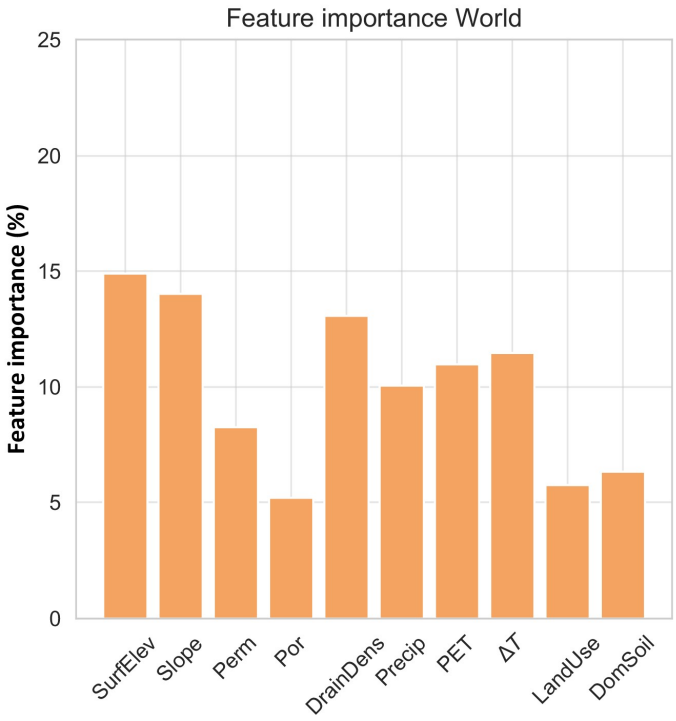

Figure B-16 Feature importance from the Global Random Forest model