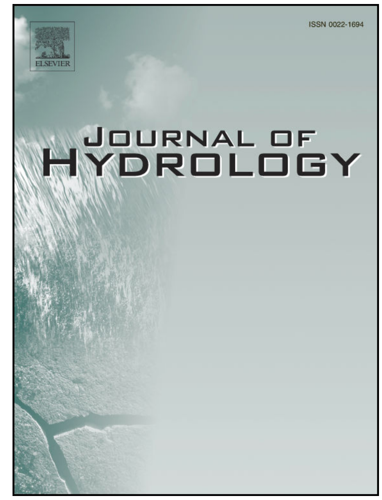
Journal Pre-proofs



Research papers

Machine learning for faster estimates of groundwater response to artificial aquifer recharge

Valdrich Jude Fernandes, Perry de Louw, Ruud Bartholomeus, Coen Ritsema

PII:	S0022-1694(24)00813-8
DOI:	https://doi.org/10.1016/j.jhydrol.2024.131418
Reference:	HYDROL 131418
To appear in:	*Journal of Hydrology*
Received Date:	23 August 2023
Revised Date:	9 April 2024
Accepted Date:	11 May 2024

Please cite this article as: Jude Fernandes, V., de Louw, P., Bartholomeus, R., Ritsema, C., Machine learning for faster estimates of groundwater response to artificial aquifer recharge, *Journal of Hydrology* (2024), doi: https://doi.org/10.1016/j.jhydrol.2024.131418

Machine learning for faster estimates of groundwater response to artificial aquifer recharge

Valdrich Jude Fernandes[1*], Perry de Louw[2,1], Ruud Bartholomeus[3,1] and Coen Ritsema[1]

1. Soil Physics and Land Management, Wageningen University & Research, Wageningen, the Netherlands
2. Deltares, Utrecht, the Netherlands
3. KWR Water Research Institute, Nieuwegein, the Netherlands

* Corresponding author (email address: valdrich.fernandes@wur.nl, postal address: P.O. Box 47, 6700 AA Wageningen, The Netherlands)

Abstract

Groundwater models are a valuable tool in optimising the decisions influencing groundwater flow. Spatially distributed models represent the groundwater level in the entire area from where essential information can be extracted, directly aiding in the decision-making process. However, these models are time-consuming, limiting the number of scenarios that can be considered. This study explores different machine learning (ML) models as faster alternatives to predict the increase in steady-state groundwater head due to artificial recharge in the unconfined aquifer while considering a wider spatial extent (832 columns x 1472 rows totalling 765 km^2) than previous ML groundwater models. We trained three ML models (encoder-decoder, U-Net, and attention U-Net) with various hypothetical artificial recharge sites (100, 300, 500, and 1000 sites) in the Baakse Beek catchment (the Netherlands), using a detailed numerical groundwater model, AMIGO. The applied recharge rate along with geo-hydrological properties from the AMIGO baseline run were used as inputs to the ML models. The properties' permutation importance indicated that all properties of the first aquifer were important to predicting the response and were included when training the ML models. All three ML models improved with additional training sites but showed limited benefits from more than 500 recharge sites. Of the three ML models, U-Net and Attention U-Net outperformed the encoder-decoder. These two models achieved Nash-Sutcliffe efficiency (NSE) of more than 0.8 when trained with 300 or more recharge sites. U-Net trained on 1000 recharge sites had the highest overall NSE of 0.95. U-Net better captures input features with highly variable spatial characteristics, such as rivers and drains which influence the maximum height of the groundwater response. The model captured the influence of the input features on the response, reproducing the response patterns across the entire catchment. Finally, we showed that the trained ML models are faster than the numerical model, predicting within 0.24 seconds (97th percentile), making it ideal for optimising decisions.

Keywords:

Managed Artificial Recharge; Machine learning; Drought mitigation; Scenario optimisation; Groundwater Response

1. Introduction

In the context of international policy frameworks like the European Water Framework Directive and Natura 2000, water management authorities have multiple targets they need to meet. The droughts of 2018-2020, which set a new benchmark in Europe (Rakovec et al., 2022), increased the urgency to take appropriate measures (Bartholomeus et al., 2023). Although the events are considered rare in the current climate, future climate change could exacerbate such events (Aalbers et al., 2023; Balting et al., 2021; Lehner et al., 2017; Pronk et al., 2021; van der Wiel et al., 2021). Even in deltas like the Netherlands droughts cause serious risks for nature, agriculture, infrastructure and drinking water availability, which resulted in drought-related policy actions like "Water and soil leading in land use planning" (Bartholomeus et al., 2023). One of the reasons for this vulnerability is the expansion of the surface drainage network and the increased exploitation of groundwater resources (Ahmadalipour et al., 2019; Bartholomeus et al., 2023; Castle et al., 2014; de Wit et al., 2022; Thatch et al., 2020; Thomas and Famiglietti, 2019; Witte et al., 2018).

The Pleistocene uplands of the Netherlands have recently faced severe rainfall deficits (Brakkee et al., 2022; Philip et al., 2020), increasing the reliance on surface and groundwater for irrigation. This has increased the strain on the limited water available for nature (van den Eertwegh et al., 2020). Long-term structural changes are identified to be more effective at reducing the strain than reactive, ad-hoc remedies during droughts. Van den Eertwegh et al. 2020 recommend increasing freshwater availability through more sustainable drainage networks, reducing groundwater abstraction, and increasing groundwater recharge.

Managed aquifer recharge (MAR) can increase freshwater availability during dryer periods by storing water surplus from the wetter periods in the subsurface (Dillon et al., 2020, 2019; Hartog and Stuyfzand, 2017). It is often categorised into infiltration, direct injection, and filtration techniques (Casanova et al., 2016); we focus on infiltration techniques that recharge the water table from infiltration basins or subsurface infiltration systems, often making them the cheapest technique. However, water managers need to identify the optimal location, recharge rate and combination of the recharge sites when designing the solution which is often done using a numerical groundwater model. These models use a set of mathematical equations to estimate the flow of water within a grid that represents the hydrological system by their characteristics, such as the aquifer's transmissivity, resistance and the surface drainage network. However, they are complex and simulating multiple scenarios for optimisation is time-consuming, limiting structured exploration and selection of potential recharge sites across an area. To facilitate the exploration of suitable recharge sites, there is a need for fast calculating tools to estimate the effect of managed aquifer recharge quickly. For such optimisation applications, a less accurate but faster option with interpretable results could be more suitable (Newman, 1996).

Such an option could be a surrogate model, which is a simplified representation of a complex, higher-order model (Wang et al., 2014). Reduced order models have been applied in groundwater modelling as surrogate models for their computational efficiency (Boyce et al., 2015; Dey and Dhar, 2020; Stanko et al., 2016; Vermeulen et al., 2004). Proper orthogonal decomposition, a common reduced-order modelling method, identifies the lower dimensional basis that captures the high-dimensional dynamics of the system. Vermeulen et al. (2004) have demonstrated its applicability in reproducing groundwater heads in a linear system. In a realistic case study, they achieved a relative mean absolute error of less than 6% while realising a 625x speed up. However, these attempts have been made for confined conditions with linear behaviour. Boyce et al. (2015) and Stanko et al. (2016) expanded this technique to unconfined aquifers, increasing the nonlinear behaviour due to the boundary conditions such as rivers. While more realistic, they are still limited to small synthetic systems with less than 200 by 200 cells. Furthermore, proper orthogonal decomposition models are limited to the location used to calculate the reduced space.

Machine learning (ML) has recently been a frequently used surrogate model as a universal function approximator. It can learn nonlinear relations in the data, which can be the results from existing numerical models. It has been used to reproduce models in fluid dynamics (Brunton et al., 2020), material science (Papadopoulos et al., 2018) and earth system models (Kim et al., 2015; Weber et al., 2019), among others. Deep learning models have been used in groundwater modelling to forecast the head at wells (Malik and Bhagwat, 2021; Müller et al., 2021; Tao et al., 2022). Asher et al. (2015) and Miro et al. (2021) recognised the lack of spatially distributed representation of groundwater surrogates. Since then, some authors have demonstrated the applicability of the convolutional encoder-decoder model, which satisfies this requisite (He et al., 2021; Mo et al., 2019; Taccari et al., 2022). However, these applications are also limited to small synthetic systems.

Artificial groundwater recharge affects the groundwater head in a large spatial area. This entire spatial extent needs to be captured by the ML model. The applicability of the above ML models at reproducing the results from a numerical groundwater model with actual subsurface properties of an aquifer has not been demonstrated yet. Furthermore, the ML model can be more specialised and represent the priorities of the optimisation challenge rather than a model reproducing all details of the system. We investigate the performance of three ML models for a catchment within the sandy uplands of the Netherlands and quantify the effect of artificial recharge for all possible locations within the area. The ML models' output is the increase in the steady-state phreatic groundwater head, henceforth groundwater response, to applied recharge sites in the Baakse Beek catchment in the Netherlands. The hydrological properties and the results from a detailed numerical model (AMIGO) are used to train the ML models. The ML model is trained on the geo-hydrological properties of the first aquifer for a wider domain size of 1472 columns by 832 rows at a 25x25 m resolution representing a 765 km^2 area. In doing so, we consider various combinations of geo-hydrological properties within the catchment and their impact on the performance of the surrogate model at predicting the steady-state groundwater head response to artificial recharge. These steps are further elaborated in the methodology and through the flow chart in Figure 1. The central questions this study aims to answer are:

1. Is the surrogate model able to reproduce the steady-state groundwater head response to artificial recharge with sufficient accuracy?
2. Which physical characteristics are required to capture the steady-state response of the groundwater head to artificial recharge in a surrogate model trained on the results of a numerical model?
3. How much training data is needed to train the surrogate model to sufficient accuracy?

In addressing these questions, this paper aims to aid future modellers in designing more accurate ML models for scenario optimisations. These questions remain relevant even through the fast advancement in artificial intelligence and ML. Multiple geo-hydrological properties represent the subsurface, but identifying the most relevant properties could help the ML model capture the relation between them and reduce overfitting. Furthermore, we want to minimize the number of slow numerical model runs. This paper compares the performance of the ML model when trained on datasets of various sizes. This offers an estimate of the number of scenarios needed to train the ML models and the effect of additional scenarios on the predicted groundwater response. Comparing three ML models with increasing complexity offers a more general view of answering the above questions and model complexity necessary to represent the relation between the recharge rate, hydrogeological properties and the groundwater response.

2. Methodology

The research methodology consists of two main parts: numerical modelling and machine learning modelling (Figure 1). The goal is to use the numerical model to simulate a baseline steady-state scenario of natural recharge and steady-state scenarios with artificial recharge at sites across the study catchment. The difference in the groundwater heads between the artificial recharge scenarios and the baseline scenario is the groundwater response to the artificial recharge. A machine learning model is trained to reproduce this response. The rate of artificial recharge (5-25 mm/day) and site size (0.01-1 km^2) are selected randomly using Latin Hypercube Sampling to represent the entire range of potential recharge sites. Orthogonal Array-based Latin Hypercube Sampling is used to select the site location, within the model extent, as it samples the location more uniformly.

The ML models are trained to reproduce the steady-state groundwater head response due to artificial recharge from the numerical groundwater model, AMIGO. These ML models are trained on training datasets of various artificial recharge realizations. Each realization contains six inputs from the AMIGO baseline run: (1) the artificial recharge rate, (2) baseline groundwater depth, (3) river stage and drain level relative to the baseline groundwater head, (4) river conductivity, (5) transmissivity of the first aquifer and (6) hydraulic resistance below the aquifer. The inputs were included based on their permutation importance in estimating three key characteristics of the steady-state groundwater head response to artificial recharge, namely the maximum, area, and total response. The ML model performance is also assessed on the same three key characteristics as they describe the most relevant properties of the response to optimize.

For steady-state simulations, the storage coefficient is zero by definition, thus not an input of the numerical model simulations, and therefore also not included in the inputs for the ML model. It should be noted, however, that in transient simulations the storage coefficient will be another system characteristic that importantly influences aquifer storage capacity to artificial recharge. Additionally, using the storage coefficient from a transient model lets us estimate the extra volume of water which can be stored by the artificial recharge, based on the simulated head differences.

Three ML models are trained using the listed inputs: encoder-decoder, U-Net and Attention U-Net, with increasing numbers of recharge sites: 100, 300, 500 and 1000. These models are designed to be increasingly complex, with the Attention U-Net having the highest number of parameters.

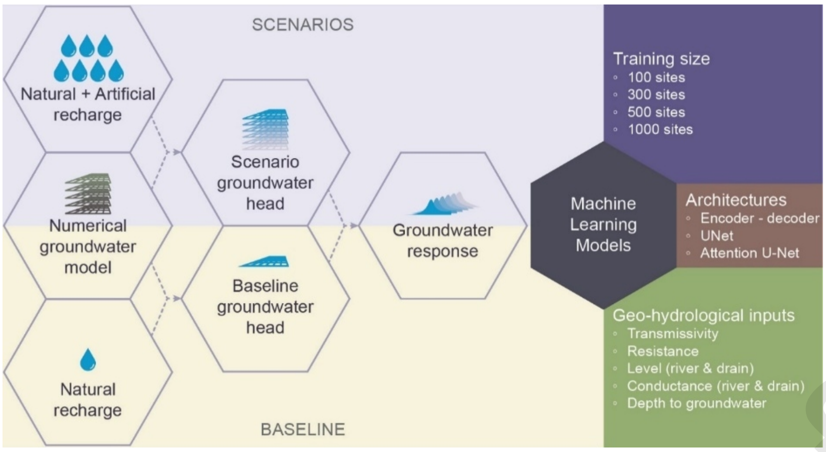


Figure 1 Steps performed before training the machine learning models to reproduce the groundwater response to additional aquifer recharge. The groundwater response is the increase in the steady-state groundwater head in the scenarios with the artificial recharge over the baseline scenario. The scenarios were simulated using the numerical groundwater model AMIGO. We compared the importance of different geo-hydrological inputs, machine learning model architectures and the number of scenarios necessary to train the model.

2.1. Numerical Modelling

The ML model is designed to reproduce the steady-state response to additional artificial recharge in the Baakse Beek Catchment east of the Netherlands, as simulated by the AMIGO numerical groundwater model. The catchment drains an area of 262.5 km^2 into the IJssel, a distributary of the river Rhine (Figure 2). This catchment is in the Netherlands' higher sandy region, insert in Figure 2, characterized by a 200m-thick sequence of Pleistocene sands intercalated with thin clay beds, which become thicker towards the west. It is mainly composed of coarse-textured glacial and beach deposits (Hijma, 2017; Sevink and Koopman, 2020), which are highly transmissive.

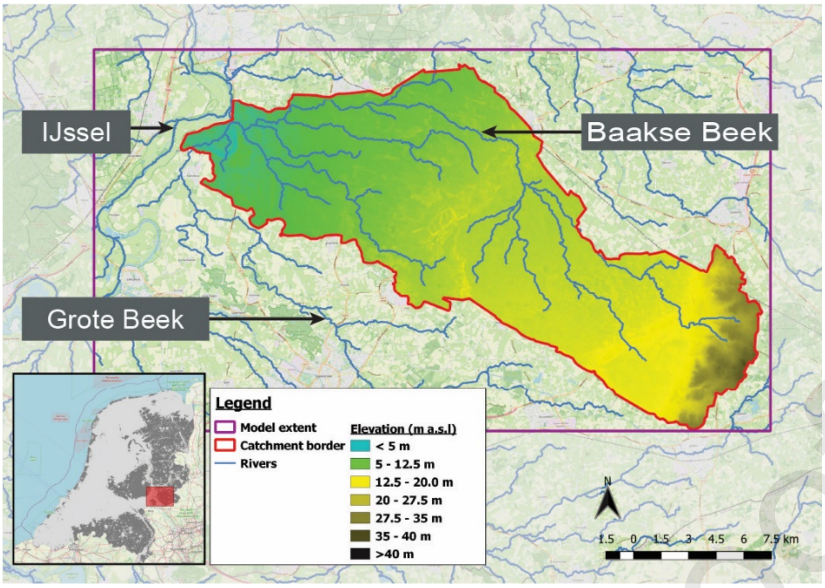


Figure 2 Map of the study area, Baakse Beek catchment, in the sandy region of eastern Netherlands (the dark grey region in insert)

The Baakse Beek catchment is represented in the spatially distributed regional groundwater model AMIGO (Actueel Model Instrument Gelderland Oost v3.1) which covers the eastern region of the province of Gelderland. It is widely used by the regional water management authority Rijn en IJssel, province of Gelderland, drinking water companies, and consultancies. The model was calibrated and validated by its maintainers (Vreugdenhil, 2021). Within Baakse Beek, the maintainers determined the modelled average low groundwater level is 5 cm higher and the average high groundwater level is 22 cm lower than the observed levels in 2008-2016.

The AMIGO model consists of 15 layers, represented by their transmissivity and the hydraulic resistance between them at a 25m resolution. The hydraulic resistance is calculated as saturated thickness divided by the vertical hydraulic conductivity of the aquifers and the resistive layer between them. This model, which includes tile drainage, ditches, streams, and extraction wells, is implemented in iMOD (Vermeulen et al., 2021) for MODFLOW-2005 (Harbaugh, 2005). In AMIGO, installed tile drainage are modelled using the DRN package in MODFLOW while ditches and streams are modelled with the RIV package. These two packages together represent the surface water network that drains the groundwater. To help the model capture the effect of the surface water network, the two packages are combined in a common input, referred to here as DRN and RIV. The AMIGO model was then cropped to a rectangle containing the Baakse Beek catchment. A fixed head boundary condition was defined along the edge of a rectangle surrounding the study catchment. The boundary is maintained at a distance of three times the leakage factor from the catchment's boundary to ensure that the boundary does not significantly influence the calculated response to recharge sites within the catchment. The groundwater head from a steady-state run with long-term temporal average natural recharge is used as the model's initial and boundary conditions.

2.2. Machine learning models

General modelling task - The results from the numerical model scenarios are used to train a machine learning (ML) model using a surrogate modelling approach. In this approach, a surrogate model (f') is used to approximate the results (y) of a complex model (f) by reproducing its outputs. However, in this study, the ML model predicts the difference between the results from a natural recharge scenario (f_o) and the artificial recharge scenario (f_s) (equation 2) rather than the complex model results directly (equation 1). This increases the relevance of the surrogate model to the scenario optimization task. Predicting the difference also reduces the output range, improving the training process for ML models.

$$f'(x^{N_x^* \times H \times W}) \approx f(x^{N_x \times H \times W}) = y^{N_y \times H \times W} \quad (1)$$

$$f'(x^{N_x^* \times H \times W}) \approx f_s(x^{N_x \times H \times W}) - f_o(x^{N_x \times H \times W}) = y^{N_y \times H \times W} \quad (2)$$

The spatially distributed models, like the AMIGO model, use N_x geohydrological features of size $H \times W$ to predict N_y outputs. The ML model aims to estimate the same results based on fewer input features (N_x^*) than the numerical model. This reduction in input features helps train a more generalised and representative ML model (Kutz and Brunton, 2022). However, the model needs minimum input features to capture all relevant relations. The numerical groundwater model requires 105 two dimensional features, while the ML models reproduce the response based on 6 input features.

Model architecture - Convolutional neural networks (CNN) (LeCun et al., 2015; Lecun et al., 1998), a popular ML model for image processing, are utilised in this study. These networks are especially suited for learning the local relations within the input features, which can influence the groundwater system in neighbouring grids. In the context of this paper, a feature is a measurable property that is input to the subsequent model layers. CNNs combine multiple layers to extract different features from the input, using trainable weight matrices (filters) that consider the surrounding cells of the cell of interest. Deeper layers in CNNs extract higher-order features, while initial layers extract elementary features. These higher-order features are crucial to capture interactions between the input features (Lerman et al., 2021). In addition to the layers with filters, CNNs also consist of convolutional, upsampling, batch normalisation (Ioffe and Szegedy, 2015), leaky ReLU (Maas et al., 2013) and dropout layers (Srivastava et al., 2014), which together enable learning nonlinear relations between the input features.

This study compares three ML models: encoder-decoder, U-NET, and attention U-NET. The three models are based on an encoder-decoder architecture. This architecture consists of encoder blocks (left block in Figure 3B) that learn the context in the input features and decoder blocks (right block in Figure 3B) that reconstruct the results from the learned context. The models differ in their encoder-decoder architecture, with variations in the number of filters.

The three encoder-decoder models share the same set of 6 input features. The inputs are two-dimensional matrices, i.e. spatially distributed values, of artificial recharge rate, aquifer transmissivity, vertical hydraulic resistance, DRN and RIV conductance, DRN and RIV stage relative to the groundwater head of the baseline run and the depth to the groundwater head of the baseline run (Figure 3A). The features are selected to represent the groundwater flow within the phreatic aquifer, whose importance is confirmed based on permutation importance. These features are passed to the first down sampling block, which generates 32 features. The number of features is doubled by subsequent down sampling blocks, up to 128 features. This limit was set to reduce the memory requirements for training the models. After the encoder block, a bottleneck (bottom Figure 3B) containing two convolution layers with 256 features was added, which improves the extent to where the recharge site influences the response.

Following the bottleneck, five decoder blocks are used to reconstruct the output with decreasing numbers of features (128, 128, 128, 64, and 32) in reverse order compared to the encoder blocks. The final up-sampling to the input dimensions was done using a convolution transpose and a convolution layer. The convolution transpose consists of 8 filters of size 4x4 with stride 2, while the convolution layer produces one feature with a 1x1 filter. Finally, a leaky ReLU activation function is applied to scale back negative values and better represent the output.

Encoder - The encoder block learns context with five down sampling blocks (left half of Figure 3B). Each block reduces the input's height and width by half using convolutional layers of 5x5 filters and a stride of two and zero padding. These layers are followed by batch normalisation, leaky ReLU activation, and a dropout rate of 10%. The batch normalisation layer normalises the features with a mean of 0 and a unit standard deviation. The dropout layer replaces a random subset of the features with 0 during each iteration of the training process, hiding those features and reducing overfitting. The leaky ReLU activation introduces non-linearity to the model by scaling negative values with a slope of 0.2. It is preferred over ReLU, which only considers positive values, to avoid the 'dying ReLU problem' due to which the model weights do not update through gradient descent. The learned features in the encoder are then passed to the decoder, which recreates the response based on the learned context.

Decoder - The decoder block increases the dimension of the features back to that of the input through five upsampling blocks (right half of Figure 3B). The three models differ in their decoders. The simplest of the models is the encoder-decoder, where each upsampling block consists of a convolutional layer followed by bilinear upsampling, leaky ReLU activation (slope 0.2), batch normalisation, and dropout (rate 10%). The convolutional layer uses 5x5 filters, a stride of 1, and zero padding.

U-Net trains on higher-level features directly from the encoder and context from the deepest part of the network through skip connections. These connections join the feature from the encoder with upsampled features from deeper parts of the network. The combined features are then processed by convolutional layers, batch normalisation, leaky ReLU activation, and dropout layers like in the encoder-decoder.

The upsampling blocks in Attention U-Net (Oktay et al., 2018) are similar to that in U-NET. However, it learns to focus on specific regions in the higher-level features using an attention block. Information is extracted from the two sources of features, upsampled contextual features and the higher-level features, using convolutional layers with 3x3 filters, a stride of 1 and zero padding. Additive importance is then calculated based on the information learnt from the two features, and non-linearity is added to the importance with ReLU activation. From these, a single importance weightage is calculated using a convolution layer with a 1x1 filter and stride one and sigmoid activation, which scales the importance between 0 and 1. The detailed features are multiplied with corresponding weights to enhance the relevance of important regions and they are then concatenated with the upsampled contextual feature. This concatenated feature is then passed through the convolutional layers, leaky ReLU activation, batch normalisation, and dropout layers, similar to the previous models. Note that the attention U-Net has half the number of filters as the other two models to stay within the memory limits.

Custom loss function - The models utilised in this study employ a type of machine learning called supervised learning (Bishop, 2006), which aims to learn a mapping between inputs and outputs based on labelled examples. Specifically, the ML model outputs are compared to groundwater response from the numerical model, AMIGO. The model's performance is evaluated using a loss function such as mean squared error (MSE) to update the model parameters through gradient descent. The training procedure is monitored by tracking the ML model's performance on a validation dataset, which is concluded when the loss does not improve through the training iterations. This validation dataset consists of the AMIGO simulation results from 100 recharge sites that the ML model was not trained on.

$$\mathcal{L}_{MSE} = \frac{1}{N^*} \Sigma (y - \hat{y})^2 \quad (3)$$

$$\mathcal{L} = (1 - \alpha) * MSE_0 + (\alpha) * MSE_r \quad (4)$$

The loss function was modified to help make it more suitable for the task. The target variable from AMIGO is sparse, consisting of multiple cells with no groundwater response, which can lead to the model primarily predicting zeros. The dying ReLU problem (Lu et al., 2020) further exacerbates this problem. To address this, the mean squared error loss ($\mathcal{L}_{MSE}$ in Eq 3) was split into two components in Eq 4: MSE for predictions where there is no response to the applied artificial recharge (MSE_0), and MSE for predictions of the response (MSE_r). In Eq 3, y and $\hat{y}$ are the groundwater response from AMIGO and the ML model respectively, while N^* is the number of input sites in the iteration. The final loss ($\mathcal{L}$) is a weighted sum of these two components, controlled by a hyperparameter α (Eq 4). This loss function offers several advantages: it balances the error between the overrepresented zeros and the response, unlike a mask, it is still sensitive to predictions away from the site, and the tuneable parameter α allows the relative importance of the two components to be adjusted to reflect the priorities of the use case.

Based on this loss, the ML model parameters were iteratively updated using the ADAM optimiser (Kingma and Ba, 2014), with a learning rate schedule. Each iteration consisted of 8 recharge sites (batch size = 8). The initial learning rate was set to 0.002, which was halved if the loss did not improve over five iterations. The training was continued until the loss did not decrease for ten consecutive iterations (Figure 3), reducing the training time compared to relying on the default reduction in the learning rate used by ADAM.

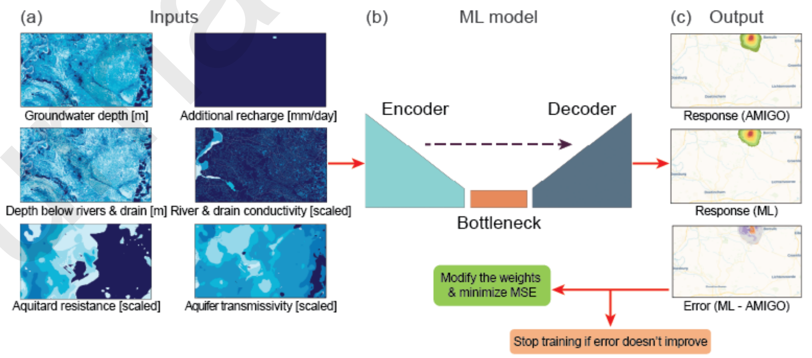


Figure 3 The training process of the machine learning (ML) models. (a) The ML model is trained on the five features from AMIGO with the recharge rate we want to predict the groundwater response. (b) The ML models are based on the encoder decoder architecture. Two variants of U-NET have skip connections from the encoder directly to the decoder represented by the dashed line. The model weights are iteratively updated during training using the ADAM optimiser to a minimise loss. This loss is based on the mean squared error (MSE) between the (c) ML model predictions and those from the numerical model AMIGO. The training iterations are concluded when the loss does not reduce on an unseen validation set for ten iterations. Basemap from OpenStreetMap-carto.

2.3. Training the model

Inputs to the ML models - The relative significance of the phreatic aquifer's properties was evaluated using the permutation importance approach (Altmann et al., 2010). This method estimates the significance by evaluating the increase in error that occurs after permuting that property. This method was used to compare the importance of the five phreatic aquifer properties and eliminate irrelevant ones. The compared properties are: transmissivity, hydraulic resistance below the aquifer, DRN and RIV conductance, DRN and RIV stage relative to the groundwater head in the baseline scenario, and the ground height relative to the groundwater head in the baseline scenario. These properties are two dimensional and need to be summarised as tabular features before estimating their importance. Four tabular features are calculated from each 2D feature which are: (1) mean, (2) minimum, and (3) maximum values where the steady-state groundwater response was more than 1 cm and the (4) average value within a 50 m radius of the site. The two definitions of the area (where the response was more than 1 cm and 50 m from the site) were included to capture the influence of the geo-hydrological properties near the recharge site and away from the site. Three key characteristics of the response were used to assess the relevance of the features and to quantify the ML model performance: the area, the maximum, and the total groundwater response (Figure 4). The area of the response is defined as the area around the recharge site with more than 1 cm of groundwater response. The maximum response is the highest, and the total response is the volume of the aquifer saturated by a response of more than 1 cm. Based on the permutation importance, all five phreatic aquifer properties are used to train the model.

Data preprocessing was performed to improve the representation of aquifers as inputs to the ML model. The 15 model layers in AMIGO are discontinuous and are often represented by thin, highly transmissive layers. For the input of the ML model, only the characteristics of the first aquifer were used. We defined the first aquifer by combining the layers until the resistance below it exceeds 200 days. This aquifer mostly consists of all 15 layers to the East and four layers towards the West. The resistance below the 15 layers was represented by the highest resistance between the layers in AMIGO (Figure 3A). The transmissivity of this aquifer is calculated based on the hydraulic conductivity and saturated thickness of the individual layers in the baseline scenario. The properties of the aquifers also exhibited right-skewed distributions with long tails, as evidenced by their interquartile ranges (Table 1). To improve the ML model's stability and performance, these properties were log-transformed and min-max scaled to 0 and 1. However, DRN and RIV stage and surface height relative to the average groundwater head were not transformed or scaled as they are linearly related to the maximum response, draining excess recharge.

Table 1 Range and interquartile range of input features from AMIGO, before scaling and log-transforming some of the inputs.

Input	Min	First quartile	Median	Third quartile	Maximum	Scaled and log-transformed
Aquifer transmissivity (m^2/day)	0.2	920	1350	1785	5034	✓
Aquitard resistance (day)	200	4610	42650	171709	171709	✓
DRN and RIV conductance (m^2/day)	0.002	7.6	10.0	18.1	5053.0	✓
DRN and RIV stage relative to the baseline groundwater head (m)	-2.7	-0.04	0.22	0.54	11.7	
Surface level relative to the baseline groundwater head (m)	0	0.9	1.2	1.7	49	

Recharge scenarios - The ML models are trained on the steady-state groundwater response to additional aquifer recharge with a certain rate applied for a certain site size, calculated by the numerical model AMIGO. Scenarios with varying applied recharge rates, site sizes, and locations were simulated to produce the data used to train the ML models. The sites were selected using Latin Hypercube Sampling (LHS) and Orthogonal Array-based Latin Hypercube Sampling (OALHS) (Sándor and András, 2004). The recharge rates applied to the topmost layer of the numerical model range from 5 mm/day to 25 mm/day, and the site sizes range from 0.01 km^2 to 1 km^2. Each site covers 16 to 1600 model cells (each model cell is 25x25 m). These ranges were selected to represent a complete range of potential recharge sites. While there are no MAR projects in the study area, there was a test site 8 km from the catchment. It was 0.58 km^2 in size and recharged 5 mm/day during the growing season (Tang et al., 2023). This site would fall within the range considered. Internationally, recharge between 250 mm and 1500 mm is applied during the growing season, which equates to 1.4 mm/day to 8.3 mm/day (de Wit et al., 2022). While some sites would fall below the range considered in this study, allowing for higher recharge rates would enable identifying the maximum potential recharge rate at the site. The recharge rate and site sizes were selected using LHS to represent the entire range.

The effect of aquifer recharge is determined by the interplay of multiple geohydrological properties that vary throughout the catchment. While the geo-hydrological properties are the same for all scenarios, we exposed the ML model to various combinations of these properties by varying where the recharge is applied within the model extent (Figure 2). The model extent covers 765 km^2, which could consist of 75969 to 720 potential recharge sites. The location of the sites was randomly selected to represent the entire model extent in datasets of 100, 300, 500 or 1000 sites. Selecting the location at random minimizes the potential for bias, ensuring better model performance for all potential recharge sites. A similar methodology is used to select locations in previous studies (He et al., 2021; Taccari et al., 2022; Tao et al., 2022). Multiple sites were simulated simultaneously while maintaining a minimum distance between adjacent sites to reduce their interaction. Simulating multiple sites limited the number of numerical model runs. We used the OALHS method to ensure that samples are more evenly spaced, even in multiple dimensions, unlike LHS. While OALHS ensures a more uniform sampling, it does not guarantee a minimum distance between adjacent points. To enforce this condition, adjacent points are separated into groups, resulting in four numerical model scenarios from each OALHS of x and y coordinates of the recharge site's centres. Considering the dimensions of the model domain and the minimum distance, 18 sites were sampled together and then split into two groups of six and two groups of three sites. Multiple OALHS were grouped to create datasets of various sizes that represented the same sample distribution.

The results of the numerical model scenario runs were split into three datasets: four training datasets, which were created through resampling (with 1000, 500, 300, and 100 sites), a validation dataset (100 sites), and a test dataset (200 sites). The OALHS samples were maintained throughout the different datasets to ensure equal representation. The recharge sites in each dataset are shown in Appendix – A. The recharge sites in the training dataset with 1000 sites cover 364 km^2 representing 47.6% of the model extent. Of this, 47.3 km^2 overlaps with the test dataset. Although some sites in the training dataset overlap with those in the testing dataset, the recharge rate and the area of the sites differ between the sites.

2.4. Analysis

The performance of the three ML models is assessed by comparing their predictions of the three key characteristics, the maximum, the area and the total response (Figure 4). The comparison uses the Nash-Sutcliffe Efficiency (NSE) metric. The NSE measures the model's ability to explain the variance in the observations and ranges from -∞ to 1, with higher values implying a better predictive ability.

While NSE describes the overall model's performance, it does not account for systematic errors. The systemic errors across the range of responses are represented in a scatter plot of the estimated key characteristics from the best ML model and AMIGO. For this, we considered scenarios with a constant recharge rate of 15mm/day over recharge sites of 1 km^2 across the entire model domain. The key characteristics are also represented as maps that reveal the interactions between the inputs and the resulting response.

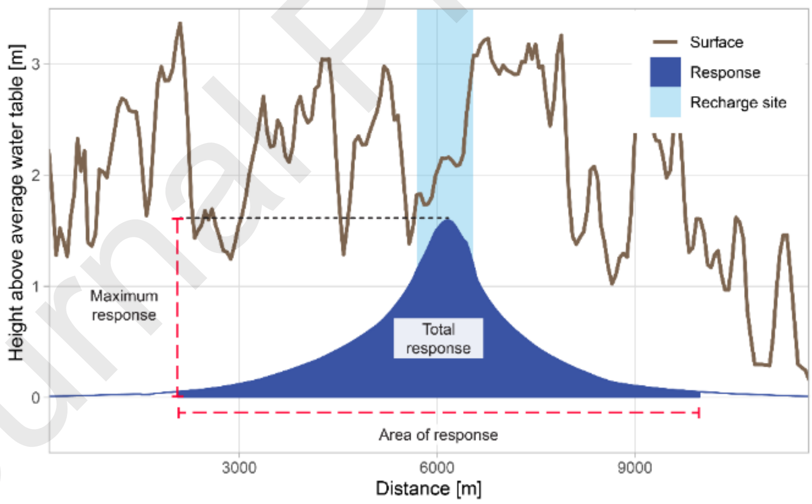


Figure 4 Cross-sectional view of a possible response of the groundwater to artificial recharge. All heights are relative to the baseline groundwater head. The increase in groundwater head (blue) is due to artificial recharge at the recharge site (light blue). The brown line represents the surface elevation relative to the baseline groundwater head. The maximum response, the area of the response and the total response are the key characteristics used to quantify the model performance. The vertical and horizontal axis are not symmetrical which exaggerates small changes in groundwater depth and response.

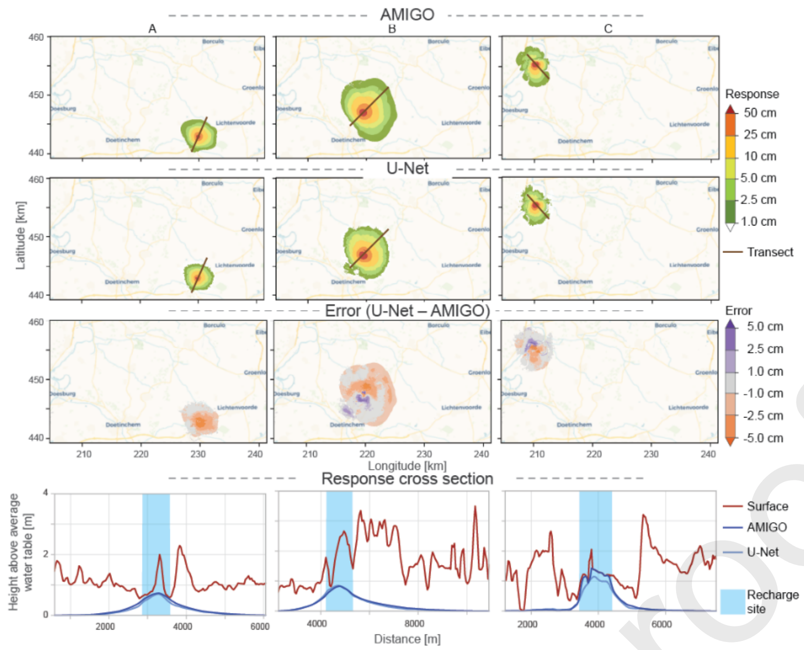


Figure 5 Map view of the response estimated for three recharge sites (A, B, C), represented in columns, by the numerical model AMIGO (top row) and by UNET model (middle row) trained on 1000 input recharge sites. The difference between the two is represented below as the error. The recharge sites are selected for their asymmetric response caused by the interaction between the groundwater and the surface water network (Groote Beek River and IJssel River). The bottom row represents the cross sectional view of the response along the transects in the maps. The vertical and horizontal axis in these cross sections are not symmetrical which exaggerates small changes in groundwater depth and response. Basemap from OpenStreetMap-carto.

To showcase the advantages of the ML model, we undertook a methodology aimed at determining the optimal location and recharge rate for sites within the catchment area. This involved simulating 7,722 recharge sites across the entire study area, each covering an area of 10 hectares. The simulation included the evaluation of eleven recharge rates ranging from 5 to 25 mm/day at 2 mm/day intervals for each site. Based on these simulations, we created a database of 84942 responses among which the optimal recharge sites can be identified. To identify these sites, we sought locations exhibiting the highest response at a low recharge rate based on the total volume of the response. Although related, this target differs from the volume of water stored in the aquifer. To estimate the extra volume of water which can be stored by the artificial recharge, we multiplied the total volume of the response by the specific yield of the phreatic aquifer. The specific yield used in AMIGO for transient simulations is 0.15 *(Vreugdenhil, 2021)* which corresponds to an aquifer composed of silt to medium sand *(Johnson, 1967)*. This aquifer material type fits the description of the Pleistocene sands in the catchment.

The assessment of these locations involved comparing their response to a constant recharge rate of 25 mm/day. Subsequently, the optimal recharge rate was discerned by identifying the minimum recharge rate that achieved more than 80% of the maximum response at each site. This comprehensive methodology allowed us to systematically analyse and pinpoint the most effective locations and recharge rates for artificial recharge within the catchment area while demonstrating the benefits of the ML model.

3. Results & Discussion

This section evaluates the performance of three ML models (encoder-decoder, U-Net and attention U-net) in predicting the steady-state groundwater head responses to artificial recharge, generated by the numerical groundwater model, AMIGO. The best performing ML model has captured the asymmetric responses to the artificial recharge (Figure 5). This asymmetry is caused by the interaction of the groundwater with the surface water network such as, with rivers and drains. The surface water network drains part of the groundwater response, hence limiting the response. Despite the added complexity, the best ML model captured this interaction, predicting the response outside the recharge site within ±10 cm. In the following sections, the performance of the ML models is further examined.

3.1. Performance of the three machine learning models

All three ML models perform well when trained on 300 or more recharge sites, indicated by high (NSE values (Figure 6). They achieved a high NSE in comparing the total response and its area despite the lower NSE for the maximum response.

Both U-Net and Attention U-Net models exhibited similar performance and consistently outperformed the encoder-decoder model. The variants of U-NET's outperformance could be due to the increased number of model parameters and the significance of the skip connections from the encoder to the decoder block in the U-Net models (Figure 3). These skip connections allow the models to capture spatially highly variable details in the input, such as DRN and RIV properties. This conclusion is further supported by the encoder-decoder model's worse performance at predicting the maximum response, as high groundwater heads are strongly influenced by the surface drainage network near them, which is better captured by the U-Net models. This effect of the surface drainage network is evident in Figure 5C, where the IJssel river (Figure 2) drains some of the groundwater, causing an asymmetric response. Similarly, the Grote Beek stream, southwest of the recharge site, causes a smaller and steeper response (Figure 5B).

Attention U-Net learns to focus on important regions within the input that help it predict the local response more accurately. Contrary to its expected better accuracy, attention U-Net does not have a significantly different NSE than U-Net. After accounting for different training sizes, the true difference in NSE between the two

ML models is between -0.06 and 0.05 (95th percentile) based on paired student's t-test. This result is counter-intuitive as the response is highly localised and the ML models could gain from focussing on selected parts of the input data. However, the attention mechanism in attention U-Net's decoder block increases the model's memory requirement, which we compensated for by halving the number of filters in the convolution layers in the model. Based on this, we can conclude that more filters greatly improve the model performance, more than the advantages of the attention layers. For models with a smaller extent, requiring less memory, it could be more beneficial to train U-Net with more filters rather than using Attention U-Net.

Furthermore, all three models improve with additional training data, particularly for the area of the response and total response (Figure 6A and Figure 6C). Specifically, the U-Net model's NSE for the predicted area increased from 0.71 to 0.96 with 1000 training sites versus 100 sites, and the NSE value for the predicted total response increased from 0.76 to 0.95 with additional training sites. However, the NSE for the predicted maximum response did not consistently improve with additional training data (Figure 6B). Additional training sites improved the performance up to 500 sites, but the predicted maximum response only marginally improved when doubling the input to 1000 sites (NSE of U-Net from 0.86 to 0.87).

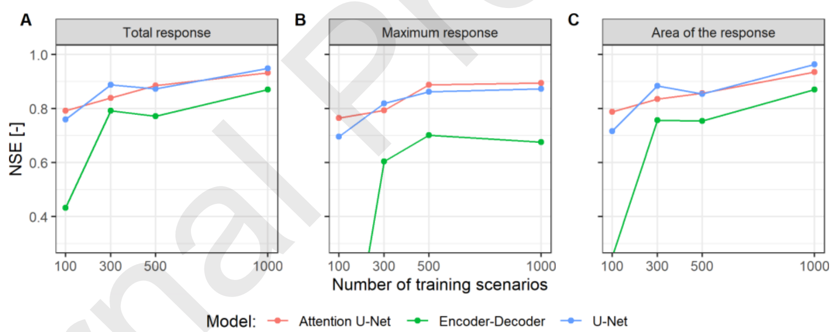


Figure 6 Nash-Sutcliffe Efficiency (NSE) of the key characteristics from the groundwater head response estimated by the machine learning models when trained with an increasing number of training sites along the x-axis. A high NSE (maximum of 1) indicated more accurate predictions. The three characteristics of the response (total, maximum and area of the response) are represented in columns (A, B and C).

Another consideration when choosing the model is the training and evaluation time. However, the training time is strongly dependent on the initial values of the parameters in the ML model and hence might not be perfectly reproduced. The initial parameters also explain the initial error that improves during the training process (Figure 7). The error does not steadily reduce during training and often fluctuates, especially early into the training. This fluctuation is likely due to a relatively high learning rate which was reduced when the training stagnated. This learning rate reduced the overall training time compared to relying on the ADAM optimiser's default learning rate. The training process seems to be slowed by the vanishing gradient problem, exacerbated by the sparse nature of the response. The encoder-decoder model trained on 100 recharge sites stagnated at this point and only predicted low responses. The model needed more than 100 sites to train further.

The relative effect of the training size on the total training time would likely be consistent in future training attempts. Additional training sites linearly increase the training time, from 70 min when trained on 100 sites to 10 hours for 1000 sites. Although it is a long time, it is 'passive time' where no human interaction is required. Each training iteration for the encoder-decoder model is shorter, but it rarely outperformed the variants of U-Net (Figure 7). Between the variants, Attention U-Net trained faster than U-Net for smaller datasets with 100 sites and 300 sites and achieved lower validation errors. This is likely due to the model's ability to learn regions to focus on through training. However, U-Net can compensate for the attention mechanism with additional training data and outperformed Attention U-Net when trained on 1000 sites.

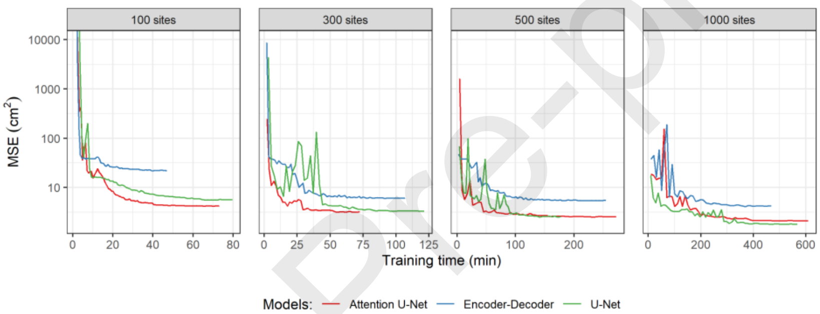


Figure 7 Validation MSE that was tracked during the training process. The MSE is calculated for an unseen set of recharge sites, validation set, different from the sites used to train the model. Additional training sites improve the final model but also increase the training time.

The evaluation time for the models ranges between 0.06 s to 0.43 s. The average evaluation time for the three models ranged between 0.09 s and 0.11 s and varied significantly between the models (Kruskal-Wallis test p-value < 0.01). However, this difference is not of practical significance, especially when compared to the average AMIGO run that took 1290 s (between 688 s and 2227 s). The slowest ML model, U-Net, could evaluate 3000 scenarios during the average time for a single scenario run in AMIGO.

3.2. Performance of the best model

The U-Net model trained on 1000 recharge sites is the best-performing ML model with the highest NSE for predicting area and total response. However, NSE does not account for systematic errors. Figure 8 shows good agreement between the U-Net and AMIGO estimates, but often U-Net underestimates the maximum groundwater response.

Figure 5C is one such scenario where U-Net underestimates the maximum response; limiting the response to the bottom of a local depression at the recharge site. The recharge rate at the site exceeds the maximum rate the groundwater can spread away from the site, leading to the groundwater head reaching the surface and seeping out through overland flow (cross-section of Figure 5C). Although such a high recharge rate is not efficient at storing water in the subsurface, the occurrence of overland flow would encourage a redesign of the recharge site. AMIGO can capture this phenomenon, but the response from U-Net does not reach the surface level. The response from U-Net is limited by the deepest surface point, resulting in a larger error at the recharge site (Figure 5C). However, U-Net underestimates the response, which still suggests that the site is inefficient at storing water and motivates redesigning the recharge site. Additionally, this error has a minor impact on the response away from the recharge site, where the response from both AMIGO and U-Net are mostly within 7.2cm of each other (99th percentile). This error is comparable to the responses in Figure 5A and Figure 5B, 9.1cm and 5.4cm respectively.

The U-Net model shows a negative bias for high values in both the total response and area of response. Specifically, the U-Net model underestimates the response of the top ten sites with the highest total response by 15% (see Figure 8A) and the area of the top ten widest response by 13%. Notably, these results are only applicable to responses more than 5cm. When including the smaller responses, up to 1cm, the bias increases to 26% (Figure 8C). Interestingly, increasing the lower limit to 10cm did not decrease the bias (13.1% vs 13.0%), indicating that U-Net underestimates the small responses and the bias increases for responses less than 5cm. Although the total response is less sensitive to the minimum limit, it still increases from 12% to 16% when considering responses less than 5cm.

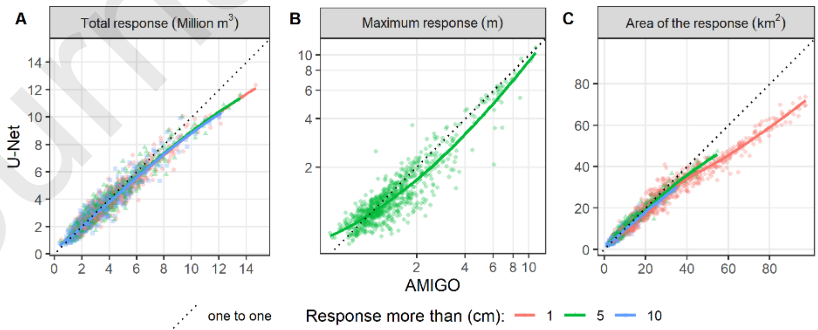


Figure 8 Scatter plot of the key characteristics of the response, estimated by U-Net vs those from the numerical model, AMIGO. These results are for recharge sites across the entire model domain, with 15mm/day recharge applied over 1 km^2. The total response and area of the response were calculated for responses of more than 1cm, 5cm and 10cm to indicate the model's accuracy at predicting smaller responses. The line is used to represent the trend in the scatter created from a local polynomial regression fitting.

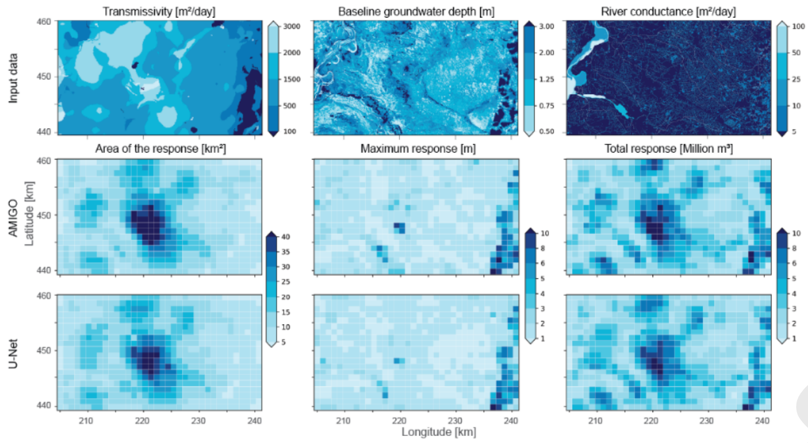


Figure 9 A comparison of the input data (top row) and the predictions of the three key characteristics (total, maximum and area of the response) from the numerical groundwater model (AMIGO, middle row) and our best machine learning model (U-Net trained on 1000 recharge sites, bottom row) for 1km² recharge sites with 15mm/day artificial recharge over the model domain.

3.3. Input features

The key response characteristics: area, maximum, and total response, depend on various hydro-geological inputs and their interaction. This interaction is evident in Figure 9, where the key characteristics are not directly related to any single hydro-geological input but a combination. U-Net could reproduce the spatial patterns of the key characteristics accurately, indicating that it has captured the effect of the interaction. Among the key characteristics, the maximum response has the most direct dependence on the groundwater depth below the ground surface and the DRN and RIV stage. These inputs limit the maximum response by draining some of the excess recharge. The recharge increases the groundwater head in the aquifer up to the drainage level. As the head increases above the drainage level, the groundwater is drained to the surface water network, depending on the head above the drain level and the drain conductance. This dependence is evident for sites at the elevated regions near the rivers. These rivers have a high conductance and hence more strongly limit the groundwater response. This dependency is also captured when estimating the importance of the inputs using the permutation importance approach (Figure 10). This approach suggests that the maximum response is significantly dependent on the average depth near the site, the maximum and minimum drain conductivity, transmissivity, and the minimum resistance.

The area of the response is the second key characteristic with a more direct relation to the input variables. The area depends on the aquifer's transmissivity, the minimum resistance between the aquifers, and the level and conductance of the surface drainage network (Figure 9) which is also reflected in the permutation importance (Figure 10). Higher transmissive aquifers allow for a faster flow of water away from the recharge site at a gentler gradient. The faster flow and a gentler gradient result in the artificial recharge providing water to a wider area. The dependency on the surface drainage network can be explained by making a comparison with groundwater abstraction. For groundwater abstractions, the equation for leakage factor is related to the area around an abstraction well where leakage occurs through the aquitard due to the pumping in the aquifer below. Higher leakage factors indicate that pumping would reduce the groundwater head in a wider area, increasing the leakage in that area. Leakage factor (λ) is the square root of the ratio of the aquifer's transmissivity (KD) and the aquitard conductance (K'/D') above the aquifer: $\lambda = \sqrt{\frac{KD}{K'/D'}}$, where K and D are the hydraulic conductivity and thickness of the layers. Phreatic aquifers do not have an overlying aquitard; for these aquifers, the properties of the surface drainage network are used instead (van der Gaast et al., 2005). Besides the effect of the aquifer transmissivity and resistance of the surface water network, the permutation importance also suggests that the area of the response depends on the minimum aquitard resistance below the aquifer (Figure 10). However, the maximum and average resistance is only significant up to a level of 5%.

The total response is the most complex and important key characteristic of the response, related to the total volume of fresh water stored using artificial recharge. It combines the other two key characteristics, i.e. the maximum and the area of the response. The total response also depends on the transmissivity and the surface drainage network properties as they affect both the maximum and the area of the response (Figure 9). Along with these inputs, the total response depends on the average groundwater depth near the site and the minimum aquitard resistance below the aquifer up to a significance level of 1% (Figure 10). Based on these results, all five features are necessary to ensure an adequate representation of the system in the ML model.

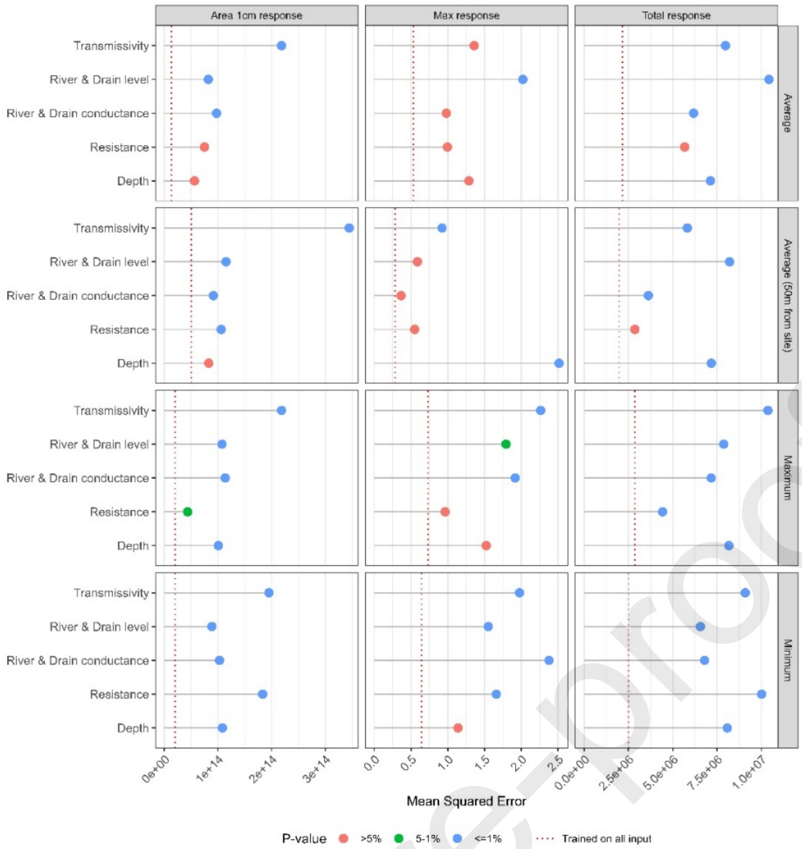


Figure 10 The permutation importance of the hydrological properties of the first aquifer. This importance is the increase in the mean squared error at predicting three performance indicators when the hydrological properties of the first aquifer are randomized. The mean, minimum, and maximum values of the property where the groundwater response was more than 1cm and the average of the property within a 50m radius of the site were used to represent the hydrological properties influencing the response at the site. The three performance indicators are (1) the area of the groundwater response, (2) the maximum response, (3) total response. P-values show the significance of the input characteristics in explaining the performance indicators. The average, maximum or minimum of the hydrological properties are important to explaining the key characteristics of the response up-to

a significance level of 0.01 and were hence included as inputs to the ML model.

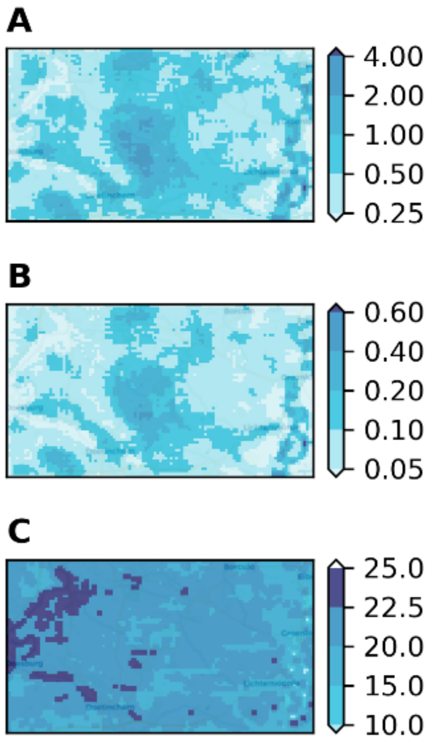


Figure 11 Optimal recharge rate for 10 ha recharge sites across the entire study area. The total volume of the response, in million m3, to recharge of 25mm/day applied in sites of 10 ha is shown in A. However, this recharge rate is often inefficient. B is the volume of water stored, in million m3, when recharging at a rate that achieves at least 80% of the response at 25 mm/day. The corresponding recharge rate, in mm/day, is in C.

3.4. Applications

The ML model's efficiency, being 3000 times faster than the numerical groundwater model, makes it suitable for various applications requiring numerous steady-state model runs. For instance, it can greatly benefit tasks like optimizing recharge rates, determining the optimal size and location of recharge sites, and comparing multiple locations rapidly. In cases where recharge volume is predetermined, such as by regulatory mandates, the ML model enables swift comparison of multiple locations. This facilitates the evaluation of various combinations of recharge rates and site areas, aiding in decision-making processes.

The bottom row of Figure 9 illustrates a notable example where the key characteristics of 720 recharge sites were compared. The ML model efficiently simulated these 720 recharge sites within 144 seconds, while the numerical model required 11 hours for the same task. To enhance the speed of the numerical model runs, each run simulated 6 equally spaced recharge sites, and five runs were executed in parallel. This comparative analysis underscores the substantial speed-ups achieved when using the ML model.

The results highlight specific regions within the catchment area, particularly the center and eastern edges, as promising potential recharge sites. Additionally, smaller regions near the northern and southwestern edges of the model domain show promise. Figure 11A depicts a similar analysis using the ML model, focusing on recharge at a rate of 25 mm/day over 10 ha sites within the model domain. The results reveal that, at this recharge rate, only the center and eastern regions exhibit a high total response. This observation suggests that different locations are more suitable at different recharge rates.

To illustrate this point, we conducted a comprehensive comparison involving the steady-state response of 7,722 recharge sites, each covering an area of 10 hectares, across 11 recharge rates ranging from 5 to 25 mm/day at 2 mm/day intervals. In total, the response from 84,942 scenarios were predicted with the ML model in 980 seconds, which would have taken the numerical model 270 days with the optimizations used to simulate 720 sites. This analysis aimed to determine the minimum recharge rate that achieves 80% of the highest total response for each site. Recharge sites located in the eastern region of the catchment achieved a high total response volume, saturating up to 4.35 million m^3 (Figure 11A), corresponding to 0.65 million m^3 water stored (Figure 11B) at the optimal recharge rate of 11 mm/day (Figure 11C). This effectiveness could be attributed to the relatively low subsurface transmissivity, resulting in a localized response to artificial recharge. Consequently, the influence of streams and ditches away from the recharge site is minimized. The high steady-state response achieved at a low recharge rate makes this region emerge as a favourable location for artificial recharge.

Conversely, the central portion of the model domain exhibits a relatively high total response of 3 million m^3 while storing 0.45 million m^3 of water. This site benefits from a higher recharge rate of 23 mm/day (Figure 11C). Given the widespread response of these recharge sites (Figure 9), they hold the potential to effectively raise the groundwater level for the entire area, thereby enhancing water availability for the broader natural environment. This underscores the strategic importance of optimizing recharge rates based on the specific characteristics of different regions to maximize the positive impact on groundwater levels and ecosystem sustainability.

This analysis can readily incorporate variations in storage coefficients across the model domain. By leveraging available data on storage coefficients, we can optimize both stored water and increases in groundwater head. While this integrated approach would enable a comprehensive assessment of the groundwater response and the alleviation of water stress on natural ecosystems and the environment it is important to note that our current study focuses on the steady-state response which is independent of the storage coefficients and hence including the coefficient is beyond the scope of this study. This focus allows us to delve deeply into the system's long-term behaviour without the added complexity of variable coefficients.

3.5. Steady-state vs transient scenarios

Steady-state scenarios depict the groundwater heads in a state of equilibrium, where the inflows balance outflows without changes in the storage within the cells. However, these scenarios assume no changes in the boundary conditions throughout the simulation, such as recharge, DRN, and RIV properties. These scenarios are thus not intended to accurately reflect temporal dynamics, such as seasonal variations in precipitation. Nevertheless, steady-state scenarios provide valuable initial estimates, particularly for evaluating the long-term effects of adaptation measures such as artificial recharge when applying a constant recharge rate. Moreover, they require less input data than transient scenarios and are faster to simulate than transient scenarios. As a result, the data for training the ML model are often available, making our technique applicable to more areas. Having fewer input data that do not change during the simulation facilitates precise attribution of the changes between scenarios to specific inputs. This study leverages the benefits of steady-state scenarios to demonstrate the applicability of the technique to optimize artificial recharge sites.

Transient scenarios have the advantage that they can offer a more detailed depiction, especially on the response of groundwater heads and storage to artificial recharge in time, by accounting for the dynamic nature of the system, based on which we can assess the effect of seasonal variability on the system. Transient scenarios also explicitly account for the changes in storage within each time step due to the additional artificial recharge or due to seepage to the surface water network. Understanding the effect of the geo-hydrological properties that affect the changes in storage could enhance the optimization of recharge site locations. Given the successful development of an ML technique to mimic steady-state conditions, as is done in the current study, the next step to develop such an approach for transient conditions is warranted. It should be noted however that successful implementation is not a given, as complexity increases. This concerns e.g. differentiation between periods of infiltration building up certain storage (autumn and winter) and storage decay during summer seasons, for which different ML approaches might be needed.

4. Conclusions

This study aims to understand the design choices for a machine learning (ML) model to predict the steady-state groundwater response to artificial recharge. It compares three state-of-the-art ML models that best reproduce the response based on an identified subset of the geo-hydrological data. The ML models were trained on the results from a pre-calibrated numerical groundwater model to reproduce the simulated response. In doing so, the response can be estimated nearly instantly and help select appropriate artificial recharge sites and optimise the sites. The ML model's performance was judged based on their performance at three key response characteristics: the maximum response, the area of the response and the total response.

Three convolutional neural networks were trained, of which U-Net and Attention U-Net could accurately reproduce the response. These models contain skip connections that enable the model to capture spatially highly variable details in the inputs, such as DRN and RIV. Additionally, both these models have similar performance suggesting that the attention mechanism does not compensate for its memory requirement. With more available memory, training a U-NET with more filters could be more beneficial than opting for Attention U-NET. Both variants of U-NET achieved a high Nash Sutcliffe Efficiency (NSE) of 0.9 when trained on the results from 500 recharge sites. Additional training sites improved the NSE to 0.96 at predicting the area of the response and the total response, while the maximum response did not show a marked improvement to additional data. However, additional data increases the computation time to generate the data and train the model, negating some of the benefits of the speed-up from the ML model. Despite the increased computation, the trained ML models could then be used to consider more scenarios, estimating the response within 0.24 s (95^{th} percentile), significantly faster than the numerical model, which took 1290 s. The slowest ML model, U-Net, could evaluate 3000 scenarios during the average time for a single scenario run in AMIGO.

Although the ML models trained in this study have a high NSE, they have their limitations. The models underestimate the maximum response in cases where groundwater levels reach the surface. Our best model is U-NET trained on 1000 sites; it limits the head to the deepest point at the recharge site (cross-section in Figure 5C). This error leads to underestimating the total response for scenarios with a high response. Despite this underestimation, the results do not impact the final recommendation that the scenario is sub-optimal and a similar response is possible with a lower recharge rate. Furthermore, the underestimation has a minor impact on the response away from the site or on the total response which is the most important characteristic to increase the water availability. Another limitation of the model is the lower accuracy in predicting small responses of less than 5cm. However, the smaller responses have a minor impact on the total response and hence should not affect the optimisation of the recharge sites.

When training similar models, future work must decide between the geo-hydrological inputs that adequately represent the groundwater system. While the groundwater head response to phreatic aquifer recharge is mostly dependent on the properties of the phreatic aquifer itself, deeper aquifers do impact the response. The deeper aquifers have a diminishing impact on the flow which we addressed by combining the numerical model layers with a low resistance between them and focussing specifically on the first aquifer. Despite the potential for enhancing the model's accuracy by incorporating the properties of deeper aquifers, the ML models trained on the properties of the first aquifer could reproduce the steady-state response. Among the properties, we identified five crucial properties based on the results of the numerical model's scenarios and Altmann's permutation importance approach: transmissivity, resistance below the phreatic aquifer, depth to the groundwater, the water level in the surface water network and the network's hydraulic conductance to flow into the aquifer. Among these inputs, the transmissivity and surface water network properties are the most important as they impact all the key characteristics of the response. Considering the importance of these inputs, future research could focus on the effect of artificial recharge on these inputs. While the effect of higher transmissivity due to higher saturated thickness is incorporated in the numerical model simulations, the higher river stages due to greater flux to the river are not incorporated. A higher river stage would reduce the river flux which would increase the response. However, incorporating this would require generating the training data using a coupled surface water – groundwater model which is beyond the scope of this research.

Fast models for specific tasks could prove an effective aid in designing good aquifer recharge sites. The speed-up could enable the water management authorities to consider many more scenarios in and around the selected catchment. The increased need for such an approach also follows from literature, e.g. from using ML-models to explain groundwater fluctuations (Sahoo et al., 2017) and the exploration of the influence of different uncertainties including future climate conditions while considering 1872 future scenarios (Miro et al., 2021). The approach could also motivate and justify the decisions to stakeholders improving support for water conservation. While this study does not demonstrate the model's performance in other regions, a similar model could best suit that region's challenges. The model in this study could serve as a starting point, and transfer learning techniques could be deployed, reducing the number of training scenarios needed and the training time.

Finally, we identified challenges when covering a larger spatial extent by the model. The larger extent increases the spatial GPU memory required when training the machine learning model. The authors limited the size of the Attention U-Net to fit in the 16 G.B. available in NVIDIA Tesla T4 GPUs. Training an Attention U-Net with more filters could make it outperform U-Net. Similarly, the adversarial loss from generative adversarial networks (GANs) could further improve the model trained, but this required training an adversarial network alongside, increasing the memory overhead in the process.

The models in this study focus on the groundwater response within the Baakse Beek catchment in the Netherlands. Future researchers could focus on training a single model for different locations, in order to investigate to what extent an ML model could be generally applicable and usable in catchments with sparse data. However, a similar extent must be maintained to ensure it can predict the entire spatial extent of the response. Furthermore, the current model is limited to steady-state scenarios, and considering the response's evolution during dryer periods could influence design choices. Groundwater heads are deeper during dryer periods, increasing the potential response to MAR. The groundwater fluctuations near the recharge site are sensitive to the storage coefficient of the surrounding aquifer which is not considered in steady-state groundwater response. The U-Net trained in this study may be extended for more complex scenarios and can be used to capture the effect of other geo-hydrological properties.

5. Acknowledgements

This research was performed within the framework of the research program AquaConnect, funded by the Dutch Research Council (NWO, grant-ID P19-45) and public and private partners of the AquaConnect consortium and coordinated by Wageningen University and Research.

6. Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT-3 for the sole purpose of improving the clarity of the text within this work. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Appendix – A: Site locations and recharge rates for all sites in the datasets

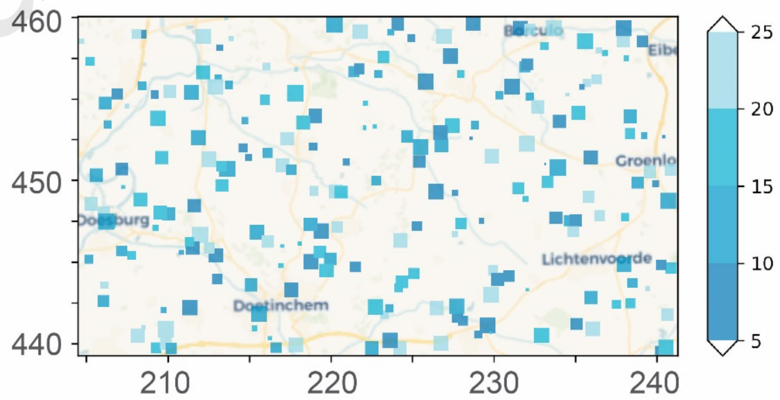


Figure A2 Recharge rate at all the recharge sites in the Testing dataset. The three ML models are compared on their performance at predicitng the response to the *recharge sites in this dataset.*

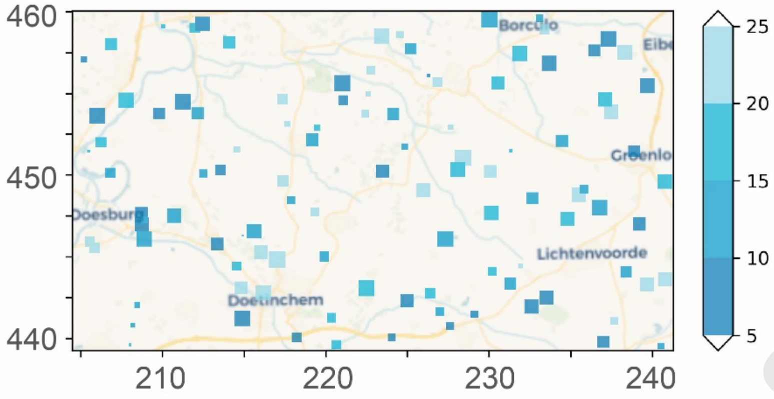


Figure A3 Recharge rate at all the recharge sites in the Validation dataset. This dataset is used to track model performance during training.

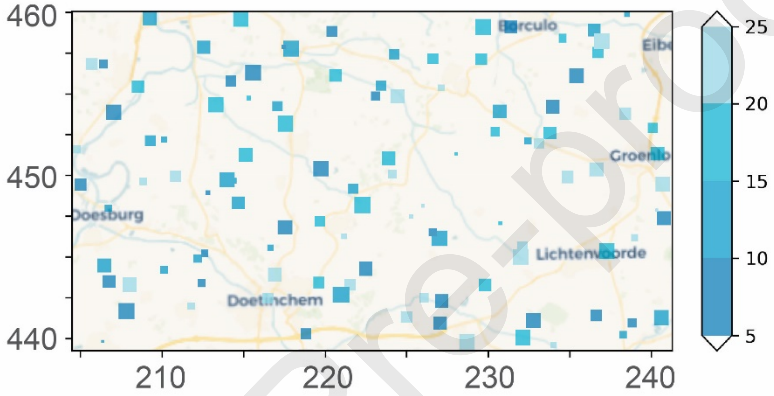


Figure A4 Recharge rate at all the recharge sites in the Training dataset with 100 sites

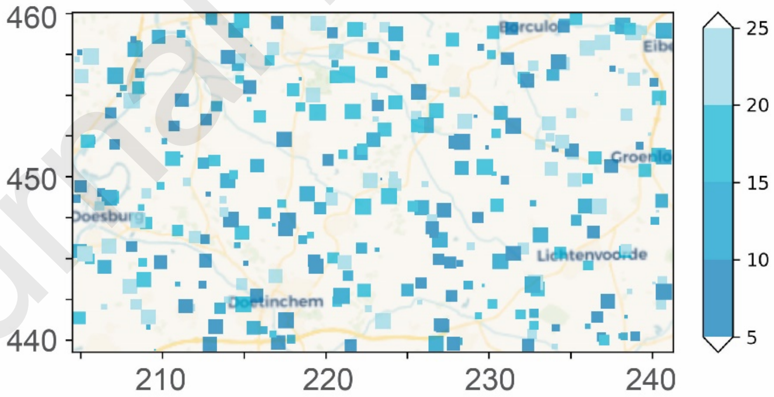


Figure A5 Recharge rate at all the recharge sites in the Training dataset with 300 sites

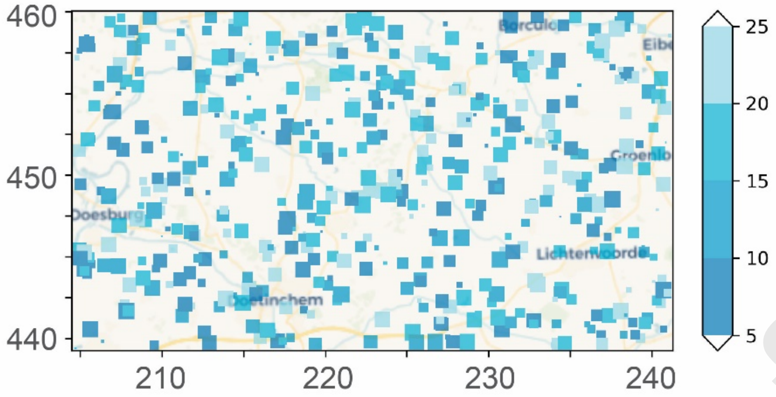


Figure A6 Recharge rate at all the recharge sites in the Train dataset with 500 sites

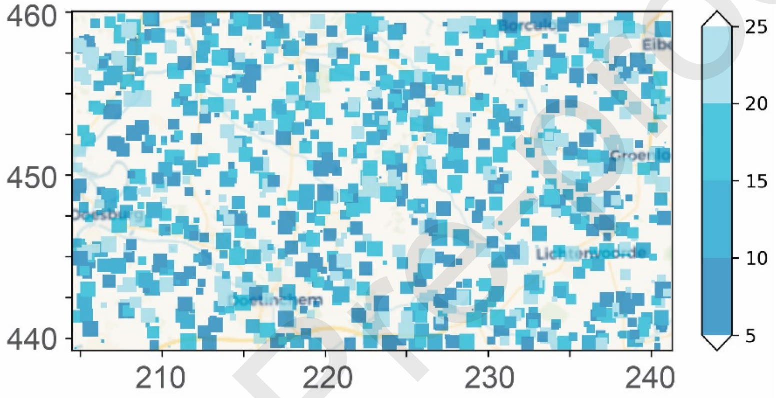


Figure A7 Recharge rate at all the recharge sites in the Train dataset with 1000 sites

7. References

Aalbers, E.E., van Meijgaard, E., Lenderink, G., de Vries, H., van den Hurk, B.J.J.M., 2023. The 2018 west-central European drought projected in a warmer climate: how much drier can it get? Nat. Hazards Earth Syst. Sci. 23, 1921–1946. https://doi.org/10.5194/nhess-23-1921-2023

Ahmadalipour, A., Moradkhani, H., Castelletti, A., Magliocca, N., 2019. Future drought risk in Africa: Integrating vulnerability, climate change, and population growth. Sci. Total Environ. 662, 672–686. https://doi.org/10.1016/j.scitotenv.2019.01.278

Altmann, A., Toloşi, L., Sander, O., Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. Bioinformatics 26, 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

Asher, M.J., Croke, B.F.W., Jakeman, A.J., Peeters, L.J.M., 2015. A review of surrogate models and their application to groundwater modeling: SURROGATES OF GROUNDWATER

MODELS. Water Resour. Res. 51, 5957–5973. https://doi.org/10.1002/2015WR016967

Balting, D.F., AghaKouchak, A., Lohmann, G., Ionita, M., 2021. Northern Hemisphere drought risk in a warming climate. Npj Clim. Atmospheric Sci. 4, 1–13. https://doi.org/10.1038/s41612-021-00218-2

Bartholomeus, R.P., Wiel, K. van der, Loon, A.F. van, Huijgevoort, M.H.J. van, Vliet, M.T.H. van, Mens, M., Geffen, S.M., Wanders, N., Pot, W., 2023. Managing water across the flood–drought spectrum: Experiences from and challenges for the Netherlands. Camb. Prisms Water 1, e2. https://doi.org/10.1017/wat.2023.4

Bishop, C.M., 2006. Pattern recognition and machine learning, Information science and statistics. Springer, New York.

Boyce, S.E., Nishikawa, T., Yeh, W.W.-G., 2015. Reduced order modeling of the Newton formulation of MODFLOW to solve unconfined groundwater flow. Adv. Water Resour. 83, 250–262. https://doi.org/10.1016/j.advwatres.2015.06.005

Brakkee, E., Van Huijgevoort, M.H.J., Bartholomeus, R.P., 2022. Improved understanding of regional groundwater drought development through time series modelling: the 2018–2019 drought in the Netherlands. Hydrol. Earth Syst. Sci. 26, 551–569. https://doi.org/10.5194/hess-26-551-2022

Brunton, S.L., Noack, B.R., Koumoutsakos, P., 2020. Machine Learning for Fluid Mechanics. Annu. Rev. Fluid Mech. 52, 477–508. https://doi.org/10.1146/annurev-fluid-010719-060214

Casanova, J., Devau, N., Pettenati, M., 2016. Managed Aquifer Recharge: An Overview of Issues and Options, in: Jakeman, A.J., Barreteau, O., Hunt, R.J., Rinaudo, J.-D., Ross, A. (Eds.), Integrated Groundwater Management: Concepts, Approaches and Challenges. Springer International Publishing, Cham, pp. 413–434. https://doi.org/10.1007/978-3-319-23576-9_16

Castle, S.L., Thomas, B.F., Reager, J.T., Rodell, M., Swenson, S.C., Famiglietti, J.S., 2014. Groundwater depletion during drought threatens future water security of the Colorado River Basin. Geophys. Res. Lett. 41, 5904–5911. https://doi.org/10.1002/2014GL061055

de Wit, J.A., van Dam, J.C., Ritsema, C.J., Bartholomeus, R.P., van den Eertwegh, G., 2022. Ontwikkeling van drainagesystemen: Water afvoeren - vasthouden - aanvullen. Stromingen Vakbl. Voor Hydrol. 28, 45–56.

de Wit, J., Ritsema, C.K., van Dam, J.C., Van Den Eertwegh, G.A.P.H., Bartholomeus, R.P., 2022. Development of subsurface drainage systems: Discharge – retention – recharge. Agric. Water Manag. 269, 107677. https://doi.org/10.1016/j.agwat.2022.107677

Dey, S., Dhar, A., 2020. On proper orthogonal decomposition (POD) based reduced-order modeling of groundwater flow through heterogeneous porous media with point source singularity. Adv. Water Resour. 144, 103703. https://doi.org/10.1016/j.advwatres.2020.103703

Dillon, P., Fernández Escalante, E., Megdal, S.B., Massmann, G., 2020. Managed Aquifer Recharge for Water Resilience. Water 12, 1846. https://doi.org/10.3390/w12071846

Dillon, P., Stuyfzand, P., Grischek, T., Lluria, M., Pyne, R.D.G., Jain, R.C., Bear, J., Schwarz, J., Wang, W., Fernandez, E., Stefan, C., Pettenati, M., van der Gun, J., Sprenger, C., Massmann, G., Scanlon, B.R., Xanke, J., Jokela, P., Zheng, Y., Rossetto, R., Shamrukh, M., Pavelic, P., Murray, E., Ross, A., Bonilla Valverde, J.P., Palma Nava, A., Ansems, N., Posavec, K., Ha, K., Martin, R., Sapiano, M., 2019. Sixty years of global progress in managed aquifer recharge. Hydrogeol. J. 27, 1–30. https://doi.org/10.1007/s10040-018-1841-z

Harbaugh, A.W., 2005. MODFLOW-2005, The U.S. Geological Survey Modular Ground-Water Model—the Ground-Water Flow Process (Techniques and Methods), Techniques and Methods.

Hartog, N., Stuyfzand, P., 2017. Water Quality Considerations on the Rise as the Use of Managed Aquifer Recharge Systems Widens. Water 9, 808. https://doi.org/10.3390/w9100808

He, T., Wang, N., Zhang, D., 2021. Theory-guided full convolutional neural network: An efficient surrogate model for inverse problems in subsurface contaminant transport. Adv. Water Resour. 157, 104051. https://doi.org/10.1016/j.advwatres.2021.104051

Hijma, M., 2017. Geology of the Dutch coast (No. 1220040- 007-ZKS- 0003).

Ioffe, S., Szegedy, C., 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.

Johnson, A.I., 1967. Specific yield: compilation of specific yields for various materials (Report No. 1662D), Water Supply Paper. Washington, D.C. https://doi.org/10.3133/wsp1662D

Kim, S., Melby, J.A., Nadal-Caraballo, N.C., Ratcliff, J.J., 2015. A time-dependent surrogate model for storm surge prediction based on an artificial neural network using high-fidelity synthetic hurricane modeling. Nat. Hazards 76, 565–585.

Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization. https://doi.org/10.48550/ARXIV.1412.6980

Kutz, J.N., Brunton, S.L., 2022. Parsimony as the ultimate regularizer for physics-informed machine learning. Nonlinear Dyn. 107, 1801–1817. https://doi.org/10.1007/s11071-021-07118-3

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-Based Learning Applied to Document Recognition. Proc. IEEE 86, 47.

Lehner, F., Coats, S., Stocker, T.F., Pendergrass, A.G., Sanderson, B.M., Raible, C.C., Smerdon, J.E., 2017. Projected drought risk in 1.5°C and 2°C warmer climates. Geophys. Res. Lett. 44, 7419–7428. https://doi.org/10.1002/2017GL074117

Lerman, S., Venuto, C., Kautz, H., Xu, C., 2021. Explaining Local, Global, And Higher-Order Interactions In Deep Learning, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Montreal, QC, Canada, pp. 1204–1213. https://doi.org/10.1109/ICCV48922.2021.00126

Lu, L., Shin, Y., Su, Y., Karniadakis, G.E., 2020. Dying ReLU and Initialization: Theory and Numerical Examples. Commun. Comput. Phys. 28, 1671–1706. https://doi.org/10.4208/cicp.OA-2020-0165

Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier Nonlinearities Improve Neural Network Acoustic Models. Presented at the International Conference on Machine Learning, JMLR: W&CP, Atlanta, Georgia, USA.

Malik, A., Bhagwat, A., 2021. Modelling groundwater level fluctuations in urban areas using artificial neural network. Groundw. Sustain. Dev. 12, 100484. https://doi.org/10.1016/j.gsd.2020.100484

Miro, M.E., Groves, D., Tincher, B., Syme, J., Tanverakul, S., Catt, D., 2021. Adaptive water management in the face of uncertainty: Integrating machine learning, groundwater modeling and robust decision making. Clim. Risk Manag. 34, 100383. https://doi.org/10.1016/j.crm.2021.100383

Mo, S., Zabaras, N., Shi, X., Wu, J., 2019. Deep Autoregressive Neural Networks for High-Dimensional Inverse Problems in

Groundwater Contaminant Source Identification. Water Resour. Res. 55, 3856–3881. https://doi.org/10.1029/2018WR024638

Müller, J., Park, J., Sahu, R., Varadharajan, C., Arora, B., Faybishenko, B., Agarwal, D., 2021. Surrogate optimization of deep neural networks for groundwater predictions. J. Glob. Optim. 81, 203–231. https://doi.org/10.1007/s10898-020-00912-0

Newman, A., 1996. Model Reduction via the Karhunen-Loeve Expansion Part I: An Exposition (Technical Research Report No. T.R. 96-32). Institute for Systems Research, University of Maryland.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention U-Net: Learning Where to Look for the Pancreas.

Papadopoulos, V., Soimiris, G., Giovanis, D.G., Papadrakakis, M., 2018. A neural network-based surrogate model for carbon nanotubes with geometric nonlinearities. Comput. Methods Appl. Mech. Eng. 328, 411–430. https://doi.org/10.1016/j.cma.2017.09.010

Philip, S.Y., Kew, S.F., Van Der Wiel, K., Wanders, N., Jan Van Oldenborgh, G., 2020. Regional differentiation in climate change induced drought trends in the Netherlands. Environ. Res. Lett. 15, 094081. https://doi.org/10.1088/1748-9326/ab97ca

Pronk, G.J., Stofberg, S.F., Van Dooren, T.C.G.W., Dingemans, M.M.L., Frijns, J., Koeman-Stein, N.E., Smeets, P.W.M.H., Bartholomeus, R.P., 2021. Increasing Water System Robustness in the Netherlands: Potential of Cross-Sectoral Water Reuse. Water Resour. Manag. 35, 3721–3735. https://doi.org/10.1007/s11269-021-02912-5

Rakovec, O., Samaniego, L., Hari, V., Markonis, Y., Moravec, V., Thober, S., Hanel, M., Kumar, R., 2022. The 2018–2020 Multi-Year Drought Sets a New Benchmark in Europe. Earths Future 10, e2021EF002394. https://doi.org/10.1029/2021EF002394

Sahoo, S., Russo, T.A., Elliott, J., Foster, I., 2017. Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S. Water Resour. Res. 53, 3878–3895. https://doi.org/10.1002/2016WR019933

Sándor, Z., András, P., 2004. Alternative sampling methods for estimating multivariate normal probabilities. J. Econom.

120, 207–234. https://doi.org/10.1016/S0304-4076(03)00212-4

Sevink, J., Koopman, S., 2020. Maximum Holocene groundwater levels and associated extension of peat in the border zone of 'Het Gooi' (the Netherlands): a reconstruction based on the study of soil transects. Neth. J. Geosci. 99, e7. https://doi.org/10.1017/njg.2020.7

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J. Mach. Learn. Res. 15, 1929–1958.

Stanko, Z.P., Boyce, S.E., Yeh, W.W.-G., 2016. Nonlinear model reduction of unconfined groundwater flow using POD and DEIM. Adv. Water Resour. 97, 130–143. https://doi.org/10.1016/j.advwatres.2016.09.005

Taccari, M.L., Nuttall, J., Chen, X., Wang, H., Minnema, B., Jimack, P.K., 2022. Attention U-Net as a surrogate model for groundwater prediction. Adv. Water Resour. 163. https://doi.org/10.1016/j.advwatres.2022.104169

Tang, D.W.S., Van der Zee, S.E.A.T.M., Narain-Ford, D.M., van den Eertwegh, G.A.P.H., Bartholomeus, R.P., 2023. Managed phreatic zone recharge for irrigation and wastewater treatment. J. Hydrol. 626, 130208. https://doi.org/10.1016/j.jhydrol.2023.130208

Tao, H., Hameed, M.M., Marhoon, H.A., Zounemat-Kermani, M., Heddam, S., Kim, S., Sulaiman, S.O., Tan, M.L., Sa'adi, Z., Mehr, A.D., Allawi, M.F., Abba, S.I., Zain, J.M., Falah, M.W., Jamei, M., Bokde, N.D., Bayatvarkeshi, M., Al-Mukhtar, M., Bhagat, S.K., Tiyasha, T., Khedher, K.M., Al-Ansari, N., Shahid, S., Yaseen, Z.M., 2022. Groundwater level prediction using machine learning models: A comprehensive review. Neurocomputing 489, 271–308. https://doi.org/10.1016/j.neucom.2022.03.014

Thatch, L.M., Gilbert, J.M., Maxwell, R.M., 2020. Integrated Hydrologic Modeling to Untangle the Impacts of Water Management During Drought. Groundwater 58, 377–391. https://doi.org/10.1111/gwat.12995

Thomas, B.F., Famiglietti, J.S., 2019. Identifying Climate-Induced Groundwater Depletion in GRACE Observations. Sci. Rep. 9, 4124. https://doi.org/10.1038/s41598-019-40155-y

van den Eertwegh, G., Bartholomeus, R., de Louw, P., Witte, F., van Dam, J., van Deijl, D., Hoefsloot, P., van Huijgevoort, M., Hunink, J., America, I., Pouwels, J., de Wit, J., 2020. Droogte in zandgebieden van Zuid-, Midden- en Oost-Nederland: Het verhaal: analyse van droogte 2018 en 2019 en tussentijdse bevindingen. KWR.

van der Gaast, J.W.J., Massop, H.Th.L., Heuvelink, G.B.M., 2005. Monitpring van verdroging; Methodische aspected van meetnetoptimalisatie (No. 1102). Alterra, Wageningen.

van der Wiel, K., Lenderink, G., de Vries, H., 2021. Physical storylines of future European drought events like 2018 based on ensemble climate modelling. Weather Clim. Extrem. 33, 100350. https://doi.org/10.1016/j.wace.2021.100350

Vermeulen, P.T.M., Heemink, A.W., Te Stroet, C.B.M., 2004. Reduced models for linear groundwater flow models using empirical orthogonal functions. Adv. Water Resour. 27, 57–69. https://doi.org/10.1016/j.advwatres.2003.09.008

Vermeulen, P.T.M., Minnema, B., Roelofsen, F.J., 2021. iMOD User Manual version 5.3. Deltares manual.

Vreugdenhil, I., 2021. Modelverbetering AMIGO 3.1 (No. D10043782:14). Arcadis Nederland B.V.

Wang, C., Duan, Q., Gong, W., Ye, A., Di, Z., Miao, C., 2014. An evaluation of adaptive surrogate modeling based optimization with two benchmark problems. Environ. Model. Softw. 60, 167–179. https://doi.org/10.1016/j.envsoft.2014.05.026

Weber, T., Corotan, A., Hutchinson, B., Kravitz, B., Link, R., 2019. Technical Note: Deep Learning for Creating Surrogate Models of Precipitation in Earth System Models (preprint). Clouds and Precipitation/Atmospheric Modelling/Troposphere/Physics (physical properties and processes). https://doi.org/10.5194/acp-2019-85

Witte, J.P.M., Runhaar, J., Bartholomeus, R.P., Fujita, Y., Hoefsloot, P., Kros, J., Mol, J., de Vries, W., 2018. De waterwijzer natuur: instrumentarium voor kwantificeren van effecten van waterbeheer en klimaat op terrestrische natuur, Stowa rapport. Stowa.

Highlights

- U-Net accurately reproduces the groundwater response to artificial recharge
- Inputs include properties of the first aquifer, drainage network, and recharge rate
- Transmissivity and surface water networks significantly impact the response

- U-Net representing a 3000-fold speed up compared to gridded groundwater model
- Minor benefits from more than 500 recharge sites for training