www.editada.org

*International Journal of Combinatorial Optimization Problems and Informatics, 15(4), Nov 2024, 7-18. ISSN: 2007-1558.*
*https://doi.org/10.61467/2007.1558.2024.v15i4.498*

# Assessment of Supervised Machine Learning Techniques for Predicting Groundwater Availability in Mexican Aquifers

*Alberto González Sánchez*[1], *Ronald Ernesto Ontiveros Capurata*[1]
[1] Instituto Mexicano de Tecnología del Agua, Coordinación de Seguridad Hídrica, Jiutepec, Morelos, México

[1] alberto_gonzalez@tlaloc.imta.mx, ronald.ontiveros@tlaloc.imta.mx

**Abstract.** Groundwater overexploitation is a global problem. In Mexico, 653 aquifers provide 39.1% of water for consumptive use. The National Water Commission manages this resource, but predicting aquifer availability is challenging, and the number of aquifers in deficit has increased. Physical models can address this issue but require extensive resources, whereas supervised learning algorithms offer a less resource-demanding alternative. This study evaluates four machine learning techniques for groundwater availability prediction: support vector machine regression, M5' model trees, random forests (RFR), and artificial neural networks. The models were trained using climatological, land use, and concession data from 1997 to 2015 and tested with data from 2018 and 2020. Random forests performed the best, showing a high correlation coefficient and low RMSE errors. The prediction accuracy for the availability state was 81.24% for 2018 and 76.79% for 2020. Thus, RFR can effectively predict short-term water availability, aiding sustainable aquifer management.

**Keywords:** aquifers overexploitation, machine learning, support vector machine regression, artificial neural networks, M5', random forests regression

Article Info
*Received Jul 15, 2024*
*Accepted Sep 29, 2024*

## 1 Introduction

Today, the excessive use of groundwater has become a significant global issue, leading to a considerable reduction in the availability of aquifers (MacAllister, 2024). Increasing demands from agriculture, industry, and urban development drive this overexploitation. Groundwater provides nearly half of all drinking water worldwide, about 40% of water for irrigated agriculture, and about one-third of water required for industry (UN, 2022). The unsustainable extraction rates are causing groundwater levels to decline in many regions, leading to adverse environmental and socio-economic impacts, such as reduced water quality, land subsidence, and loss of ecosystem services (UN, 2023). According to the Food and Agriculture Organization (FAO), over 20% of the world's aquifers are being overexploited (FAO, 2011). This problem is further exacerbated by climate change, which alters precipitation patterns and recharge rates, making the sustainable management of groundwater resources even more challenging. Therefore, it is imperative to adopt comprehensive strategies for sustainable groundwater management, incorporating technological advancements, regulatory frameworks, and collaborative efforts across sectors to mitigate the adverse effects of overexploitation and ensure the long-term availability of this crucial resource.

In Mexico, there are 653 aquifers, contributing 39.1% of the volume destined for consumptive uses, with agriculture (60%) and human consumption (14.4%) as the principal use (CONAGUA, 2018a). In this country, The National Water Commission (CONAGUA) has been striving for organized use of this resource. Since 2001, CONAGUA has aimed to use this resource efficiently by periodically estimating an annual average groundwater availability (AGWA) per aquifer, considering the concessioned volume, recharge, and other variables, following the official norm NOM-011-CONAGUA-2000 (CONAGUA, 2000). There are five versions of the AGWA values of these 653 aquifers: 2010-2011, 2013, 2015, 2018, and 2020 (CONAGUA, 2009, 2010a, 2010b, 2010c, 2011a, 2011b, 2011c, 2013, 2015, 2018b, 2020). Historically, these publications have shown a reduction in the amount of water available for extraction and an increase the number of overexploited aquifers (Table 1). Sustainable use of water resources and a more balanced allocation of groundwater extraction permits require a complex analysis considering geographical, environmental, and climatological elements to estimate AGWA values accurately.

**Table 1.** Aquifers in deficit and their average availability according to official publications.

| DOF[1] publication date | Number of aquifers | | Total mean availability (%) |
|---|---|---|---|
| | In deficit | With availability | |
| 07/08/2010, 08/16/2010, 01/25/2011, 12/14/2011[2] | 174 | 479 | 14.51 |
| 12/20/2013 | 193 | 460 | 12.50 |
| 04/20/2015 | 203 | 450 | 11.46 |
| 01/04/2018 | 245 | 408 | -2.46 |
| 09/17/2020 | 275 | 378 | -12.01 |

However, aquifers have a dynamic nature that is difficult to model; they respond to changes in land use and cover, climate, recharge volume, and extraction (X. Wang et al., 2018). Predicting aquifer recharge is complicated since it cannot be measured directly (Crosbie et al., 2015; Gao et al., 2014) (Crosbie et al., 2015). One method to detect aquifer depletion is through physical simulation models, which require a large amount of information and are expensive because they depend on the direct measurement of field variables for their calibration and validation (Coulibaly et al., 2001). An alternative to detect aquifer depletion is the machine learning technique (ML), which can build models from previously labeled records (Han & Kamber, 2006). These models identify trends without deep knowledge of the underlying attributes used in physical groundwater flow models (Steyn, 2018). Various ML algorithms have been used around the world to address the problem of excessive groundwater exploitation (Uc-Castillo et al., 2023), for example, artificial neural networks (ANN) (Daliakopoulos et al., 2005), random forests (RF), and support vector machines (SVM) (Kanyama et al., 2020) Despite these examples, studies in Mexico are scarce. In this context, this work evaluates the use of four ML algorithms (M5', RF, RNA, and SVM) to predict water availability in Mexican aquifers.

## 2 Methodology

### 2.1 Construction of the Learning Set

The learning set was constructed starting from official data sources with historical information from 1997 to 2021 and variables affecting groundwater availability, such as temperature, precipitation, land use, and the distribution of types of groundwater extraction permits (Table 2). The response variable is the AGWA values, measured in cubic hectometers ($hm^3$), obtained from CONAGUA's periodic publications (Table 1). The AGWA values per aquifer for years prior to 2011 were estimated using information from the data sources (DS) (Table 2), based on the dates of granting of the aquifer extraction permits. This approach ensured a historical range comparable to that of the predictor variables.

**Table 2.** Data sources are used to obtain predictive attribute values.

| Data sources (DS) (identifier, subject, and format) | Description |
|---|---|
| DS1: Land use and vegetation<br>Format: vector **(**shapefile) | Vector layers with the classification of land use and vegetation. INEGI[3] land use series I-VII (1997, 2001, 2005, 2010, 2013, 2016 and 2021).<br>https://www.inegi.org.mx/temas/usosuelo/#descargas |
| DS2: Climate (temperature, precipitation, and evapotranspiration)<br>Format: raster | Raster layers with annual averages of temperatures, precipitation, and potential evapotranspiration (1997 to 2021)<br>https://www.globalclimatemonitor.org/ |
| DS3: Extraction permits and its annexes<br>Format: tabular (CSV) | The status of the extraction permits in 2019 includes aquifer, type of use (agricultural, industrial, urban), volume covered, and granting date.<br>https://datos.gob.mx/busca/dataset/concesiones-asignaciones-permisos-otorgados-y-registros-de-obras-situadas-en-zonas-de-libre-alu |

[1] Diario Oficial de la Federación (DOF): The official journal of the Mexican Constitutional Government responsible for publishing laws, regulations, agreements, and other acts issued by the powers of the Federation.

[2] The first publication of the AGWA values was made in partial installments.

[3] INEGI stands for "Instituto Nacional de Estadística y Geografía", the National Institute of Statistics and Geography in Mexico. It conducts censuses and produces produces demographic, social and economic indicators about Mexican society and economy.

The collected data underwent a cleaning and integration process. Firstly, the consistency of the keys was verified to ensure their homogeneity and ability to relate the different data sources involved. Erroneous and atypical data were corrected using the MySQL database manager. Integration was carried out considering the historical compatibility among the records, selecting representative years corresponding to the official AGWA publications and 2010 to data obtained between 2010-2011 (see Table 1). Then, a primary dataset was integrated with the attributes year, aquifer identifier, recharge volume, and AGWA volume (attribute to predict). The attribute year was used to relate this dataset to land use, matching it with the INEGI series closest in time (DS1). Thus, 2010 was associated with series IV (published between 2007-2010), 2013 with series V (published in 2013), 2015 with series VI (2016), and 2018 and 2020 with series VII (2021). An approximation of the AGWA values was obtained from the DS3 data by accumulating the extraction volume of the permits per aquifer registered before 2010, thus obtaining a high number of learning records. These volumes were associated with previous land use series; consequently, estimates of AGWA values were made for 1997 (Series I), 2001 (Series II), and 2005 (Series III). During this period, information was missing for eight aquifers (1% of the total), so they were eliminated from the entire learning set, adjusting the analysis to 645 aquifers.

The climate variables (temperature, precipitation, and evapotranspiration) were obtained from DS2 (see Table 2). Temperature and precipitation are the primary variables used in similar studies (Uc-Castillo et al., 2023). The information was extracted using intersection spatial and group statistics tools of the QGIS software using the vector layer of the aquifers (CONAGUA, 2023) and climate variables. For example, Figure 1 shows the extraction of the minimum temperature data, where the polygon of the aquifer 1001 "Valle de Santiaguillo" and the circumscribed cells of the raster layer are shown. These values were weighted by the proportion of the surface area occupied by each cell, yielding an approximate value per aquifer for each available year.

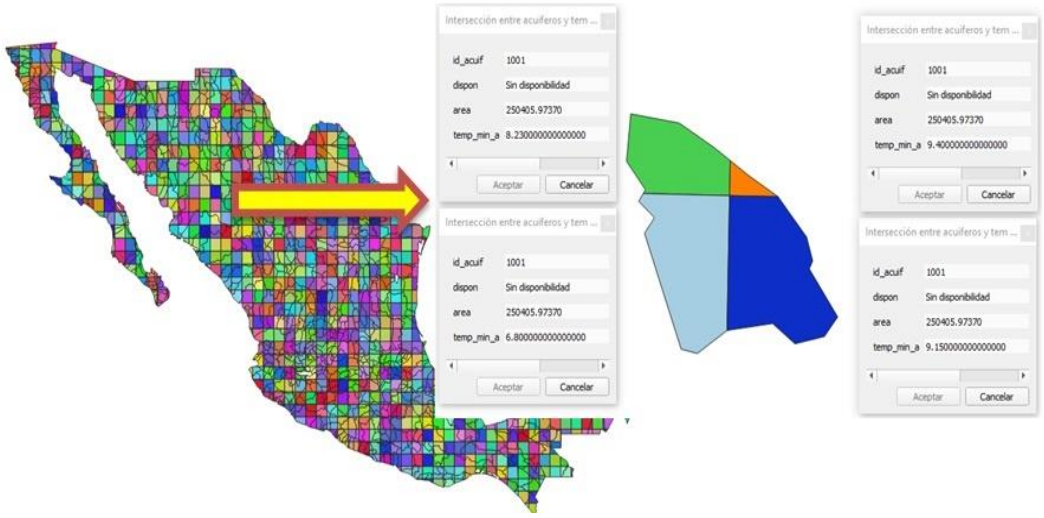


**Fig. 1.** Extraction of temperature data using the intersection of the raster with the aquifers' vector layer (QGIS).

The representative information for each year was obtained from the average of the last three years (including the reference year). The results were integrated as a training set with 5160 records (eight years for 645 aquifers) and 23 attributes (including the attribute to be predicted). The selected attributes were the following:

a) Year representative of the period to which the data in the record pertain (1997, 2001, 2005, 2010, 2013, 2015, 2018, and 2020) [YEAR].
b) Official aquifer identifier. A code number consisting of four digits, two for the State where it is located and two as a consecutive number. [ID_AQUIF].
c) Climate attributes. Average temperature (°C), precipitation ($hm^3$), and average potential evapotranspiration (mm) that occurred in the aquifer in the last three years [TEMP, PRECIP, ET].
d) Land use and vegetation. Percentage of each type of land use that the aquifer reported in the year closest to the representative one, classified as agricultural, human settlements, forest, water bodies, jungle, vegetation, and other types of coverage [S_AGR, S_SETTL, S_FOREST, S_WATER, S_JUNG, S_VEG, S_OTHER].

e) Extraction permits by distribution type. Percentage of volume authorized for extraction in the representative year classified by type of use: aquaculture, agriculture, agroindustrial, commerce, domestic, industrial, livestock, urban, services, and other uses. [R_AQUA, R_AGRI, R_AGROIND, R_COM, R_DOM, R_IND, R_LIV, R_URB, R_SERV, R_OTHER].
f) Volume available for extraction in the aquifer ($hm^3$). Variable to be predicted [AGWA].

## 2.2 Machine learning algorithms

A set of four ML algorithms commonly used for numerical prediction was selected, specifically with the implementations programmed in the Weka data mining suite (Frank et al., 2006). The following sections describe each algorithm and its parameterization for this work.

**1) Linear regression model trees M5'.** The M5' algorithm is based on a decision tree built from a recursive algorithm, making routing decisions in nodes based on the attribute values. At the end of routing, each leaf node allows the value of an instance to be obtained through a linear regression model (Gonzalez-Sanchez et al., 2014) but also can generate a numerical value; both options were tested in this work. The option of pruning the tree was also considered, generating four possible combinations: model trees with pruning, model trees without pruning, constant value trees with pruning, and constant value trees without pruning. A minimum of 2 objects were left on each leaf node.

**2) Random Forests.** Random Forest Regression (RFR) is based on the bagging method and random subspaces (Ganesh et al., 2021). The algorithm starts by generating K sets obtained by randomly drawing examples with replacements from the learning set, and each set is used to create a regression tree. In the process of constructing each tree, each partition is the product of considering a small set of input variables at random (L. Wang et al., 2016), choosing to divide the variable with the lowest Gini index. For the regression task, the result is the average estimate of the K random trees in the forest. According to (Probst et al., 2019), the most relevant hyperparameters are the number of candidate variables per partition ($m$) and the number of trees (K). The same authors suggest a value of $m=p/3$ for regression problems (where $p$ is the number of predictor attributes, 23 in this case), while Weka uses *int (log2(p) + 1))*. In this work, both options (5 and 8) were considered. These authors also suggest a value of 500 or 1000 for the number of trees. In addition, Weka allows you to specify the maximum depth of the trees using 5, 10, and without limits in this case. Therefore, for this technique, 12 combinations of parameters were validated: m ={5,8}, K ={500,1000}, and a tree depth = {5,10, unlimited}.

**3) Support Vector Machines Regression.** Support Vector Machines Regression (SVMR) belongs to a group of supervised statistical learning algorithms. In its simplest form, the objective of the technique is to obtain a linear function $f(x) = \langle w, x \rangle + b$ with $w \in R^N$ and $b \in R$ for a training set $\{(x_1, y_1), \ldots, (x_m, y_m)\}$. At most, the function f(x) should have a deviation $\varepsilon$ from the current values $y_1$ and, simultaneously, be as flat as possible. "Flatness" can be obtained with a small value for $w$. The optimization problem can be written as shown in (1) (Vapnik et al., 1997):

$$\text{Minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) \tag{1}$$

$$\text{Subject to} \quad \begin{cases} y_i - \langle w_i, x_i \rangle - b \leq \varepsilon + \xi_i \\ \langle w_i, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \tag{2}$$

Where $\xi_i$ and $\xi_i^*$ are introduced as slack variables for infeasible constraints. $C$ is the regularization parameter and determines the number of accepted deviations greater than ε. If it is impossible to separate the set of examples with a linear function, the transformation of the original space is used using a non-linear function called kernel. For this work, the version of SVMR implemented in Weka is used, which applies an improved version of the sequential minimal optimization learning algorithm (Shevade et al., 2000) with $C$=1 and a polynomial kernel of degree 1.

**4) Artificial Neural Networks (ANN).** Artificial Neural Networks are divided into an input layer, an output layer, and one or more hidden layers. The input layer consists of neurons that receive signals or data from the environment (input attributes). The hidden layer provides degrees of freedom to present more complex features. The output layer comprises neurons that provide the neural network's response. ANNs are frequently used to classify and predict historical series models (Maimon & Rokach, 2009). Different ways exist to interconnect neurons in a neural network (topology). The most common topology and training scheme, multilayer perceptron trained by backpropagation, was used for this work. For the hidden layer, combinations of 5, 10, and 15 neurons were tested, with 1000, 5000, and 10000 training cycles with decay, both parameters used in similar works carried out previously (Almuhaylan et al., 2020).

### 2.3 Evaluation

The algorithms were evaluated using the percentage-split technique, so the learning set was divided into training and testing sets. The first subset was integrated with information from the first six years (1997, 2001, 2005, 2009, 2013, and 2015), using 3870 samples (75% of the available records). The test subset was integrated with information from the last 2 years (2018 and 2020), representing 25% of the remaining records. As these are numerical prediction models, the evaluation of the test set was carried out with the metrics of correlation coefficient (r), root mean square error (RMSE), and relative square error (RRSE), which are described in Table 3 (Gonzalez-Sanchez et al., 2014).

**Table 3 .** Metrics for model evaluation. $n$ is the total number of observations; $y_i$ the real value of observation $i$; $\hat{y}_i$does the model estimate the value; $\bar{\hat{y}}$ is the mean of the model's estimates; $\bar{y}$ is the mean of observations and $r_i = y_i - \hat{y}_i$.

| Metrics | Units | Calculation |
|---|---|---|
| ***r*** | (adim) | $\frac{\sum_{i=1}^{n}(y_i - \bar{y}) - (\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}$ |
| ***RMSE*** | (same as actual value) | $\sqrt{\frac{\sum_{i=1}^{n} r_i^2}{n}}$ |
| ***RRSE*** | % | $\sqrt{\frac{\sum_{i=1}^{n} r_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} x100$ |

## 3 Results

This section shows the results of evaluating the algorithms for the test set (years 2018 and 2020). Results are presented for each technique, and an overall comparison between all techniques is made.

### 3.1 Results by algorithm

Table 4 shows the results of the M5' algorithm. The best values of each metric are highlighted in bold. Although model trees with pruning had a higher *r*, the lowest errors were obtained with a regression tree without pruning.

**Table 4 .** Metric values obtained by M5' on the testing set

| Tree type M5' | Pruning | Metrics | | |
|---|---|---|---|---|
| | | *r* | RMSE | RRSE |
| Model tree | Yes | **0.949** | 69,405 | 39.86% |
| | No | 0.945 | 74,502 | 42.78% |
| Regression tree | Yes | 0.000 | 174,141 | 100.00% |
| | No | 0.939 | **64,848** | **37.24%** |

Table 5 shows the results of the RFR algorithm. The lowest RMSE and RRSE were obtained with a forest of 1000 trees, with m=5 and no depth limit.

**Table 5 .** Metric values obtained by Random-Forest on the testing set

| Forest size (trees) | Depth limit | Variables (*m*) | Metrics | | |
|---|---|---|---|---|---|
| | | | *r* | RMSE | RRSE |
| 500 | 10 | 5 | 0.959 | 52,675 | 30.25% |
| | | 8 | 0.960 | 54,867 | 31.51% |
| | twenty | 5 | 0.958 | 53,128 | 30.51% |
| | | 8 | 0.960 | 54,752 | 31.44% |
| | Without limiting | 5 | 0.958 | 53,076 | 30.48% |
| | | 8 | 0.960 | 54,752 | 31.44% |
| 1000 | 10 | 5 | **0.961** | 50,938 | 29.25% |
| | | 8 | 0.960 | 54,582 | 31.34% |
| | twenty | 5 | **0.961** | 50,750 | 29.14% |
| | | 8 | **0.961** | 54,165 | 31.10% |
| | Without limiting | 5 | **0.961** | **50,681** | **29.10%** |
| | | 8 | **0.961** | 54,126 | 31.08% |

Table 6 shows the results of the ANNs. The best values for the error metrics were found with 1000 training cycles and 10 neurons in the hidden layer.

**Table 6 .** Metric values obtained by the RNAs on the testing set

| Training cycles | Neurons in the hidden layer | Metrics | | |
|---|---|---|---|---|
| | | *r* | RMSE | RRSE |
| 1000 | 5 | 0.928 | 91.57 | 52.58% |
| | 10 | **0.930** | **89.63** | **51.47%** |
| | 15 | 0.924 | 94.95 | 54.52% |
| 5000 | 5 | 0.921 | 103.14 | 59.23% |
| | 10 | 0.921 | 102.86 | 59.07% |
| | 15 | 0.916 | 103.47 | 59.42% |
| 10000 | 5 | 0.919 | 104.84 | 60.20% |
| | 10 | 0.919 | 104.71 | 60.13% |
| | 15 | 0.915 | 104.21 | 59.84% |

Finally, the SVMR technique was evaluated using the previously specified parameters. In this case, it was a single result, obtaining r=0.920, with an RMSE =104.675 and RRSE =60.11%.

## 3.2 Comparison between algorithms

Table 7 summarizes the best results for all the algorithms under analysis. It is observed that the RFR algorithm achieved the highest value for r and the lowest values for RMSE and RRSE. However, it should be considered that learning algorithms have varying degrees of sensitivity to their parameter values. For example, SVM is more sensitive than RF (Fang et al., 2020), and the efficiency of ANNs depends mainly on their topology and learning cycles (Haykin, 1999). Techniques like grid search, random search, or Bayesian optimization can help find the optimal hyperparameters for each algorithm, leading to fairer comparisons.

**Table 7 .** Results for evaluation metrics (all algorithms)

| Algorithm | Metrics | | |
|---|---|---|---|
| | *r* | RMSE | RRSE |
| M5' | 0.949 | 69,405 | 39.86% |
| *RFR* | **0.961** | **50,681** | **29.10%** |
| RNA | 0.930 | 89,630 | 51.47% |
| *SVMR* | 0.920 | 104,675 | 60.11% |

Once the algorithm that produces the best results has been determined, a more specific analysis can be done. Thus, the scatter plots in Figures 2 and 3 show the fit obtained by RFR in each year of the testing set.

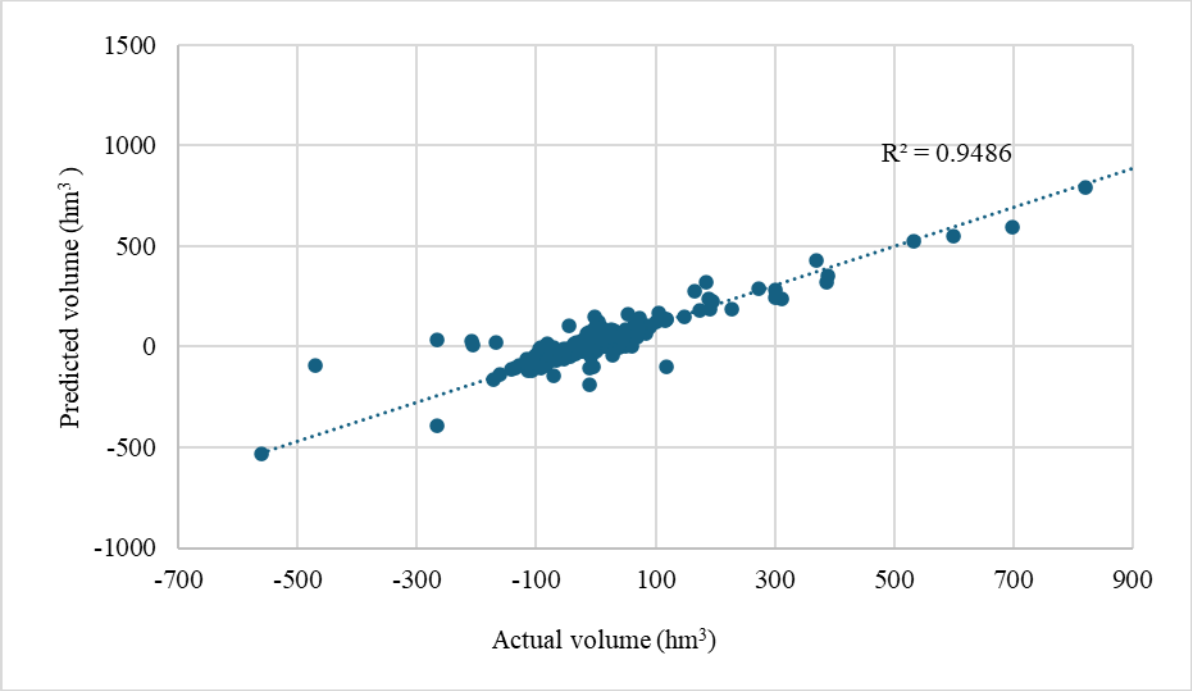


**Fig. 2 .** Actual versus RFR-estimated available volume for all aquifers (2018)

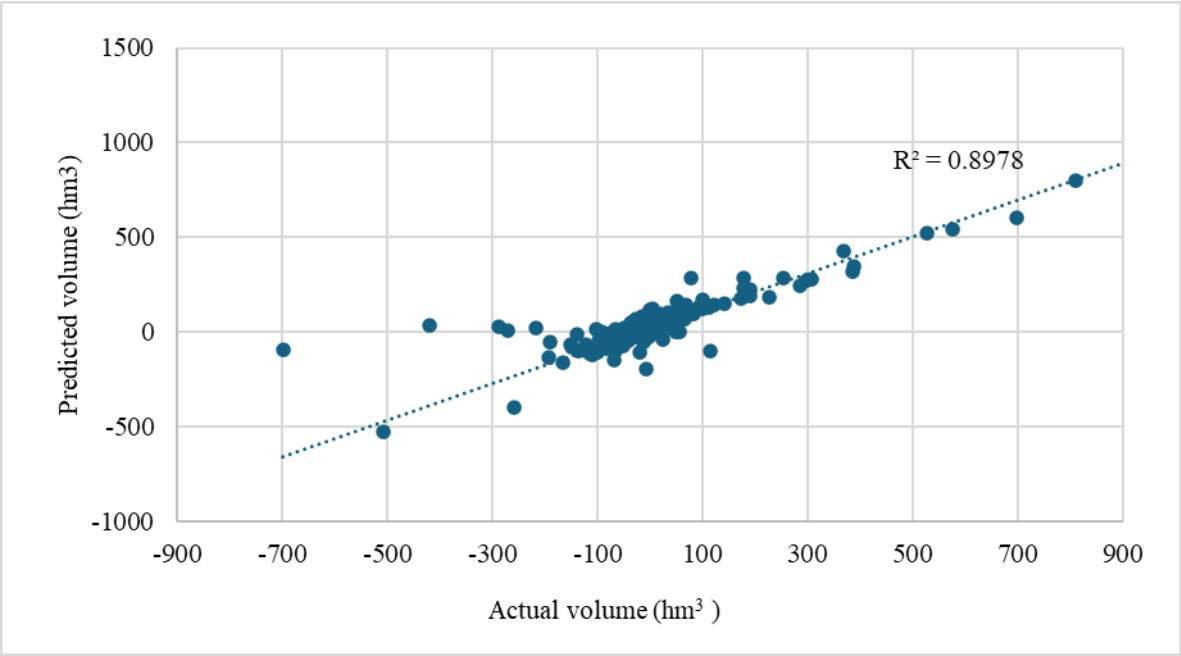


**Fig. 3.** Actual versus RFR-estimated available volume for all aquifers (2020)

Due to the large number of aquifers, a frequency analysis can improve the visualization of errors made by the algorithm. Figures 4 and 5 show the histogram of aquifers by available volume range for each year, overlaying the calculated frequency with the volume estimated by RFR. The comparison shows that the prediction has a shape that is very similar to the probability distribution of the actual AGWA value. However, it is also observed that the algorithm underestimates the number of aquifers with availability between the range of -50 to 0 $hm^3$ and overestimates in the range of 0 to 50 $hm^3$. It is consistent across the two years present in the testing set. It is easy to see a tendency for aquifers to move towards a state of availability below 0 (deficit), but the algorithm cannot identify all cases.

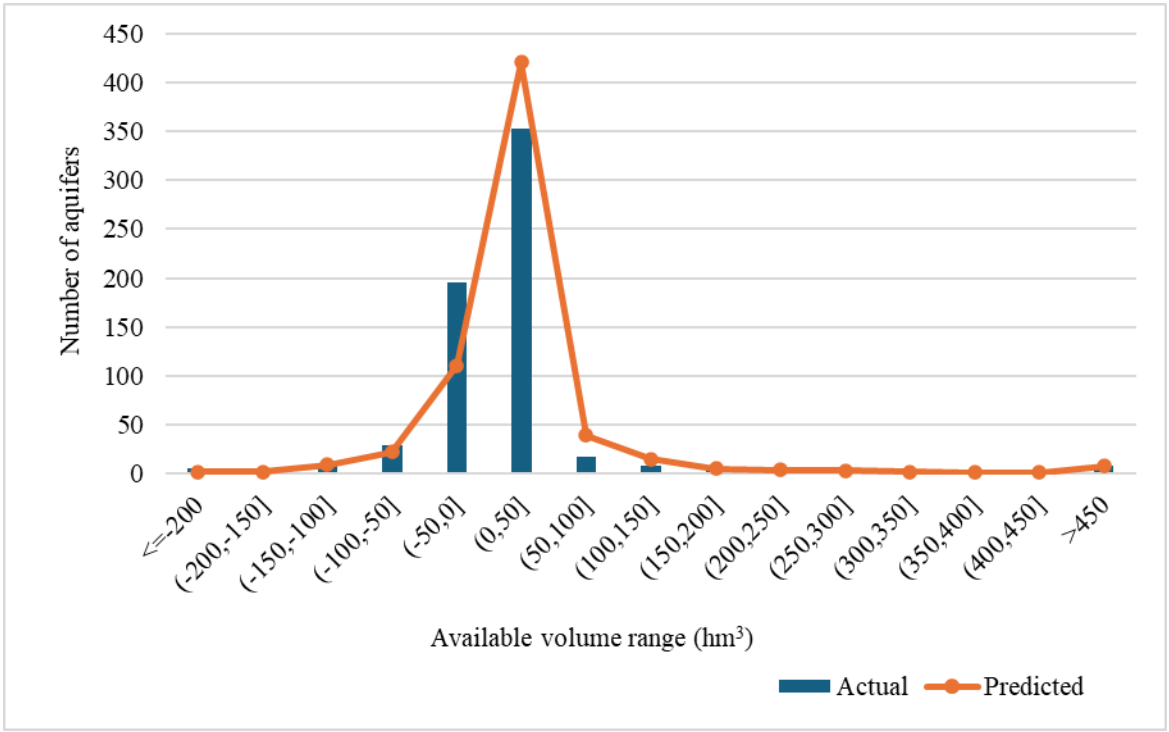


**Fig. 4 .** Histogram of available volume ranges of aquifers and RFR estimates (2018)

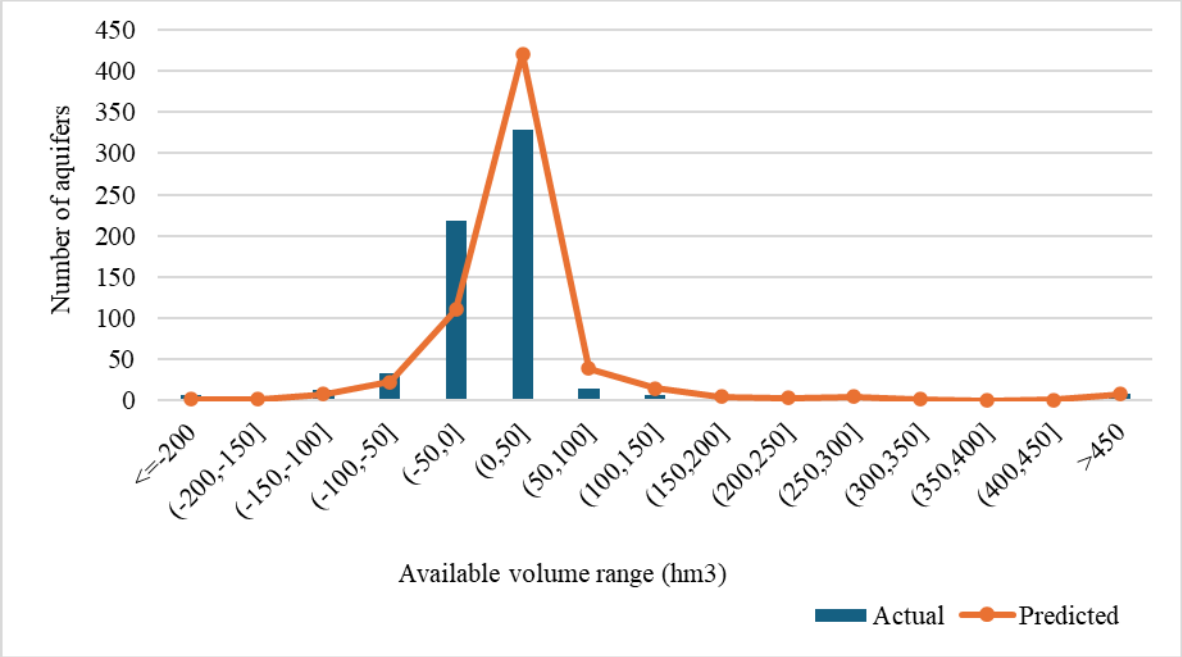


**Fig. 5.** Histogram of available volume ranges of aquifers and RFR estimates (2020)

Finally, the coincidences between the prediction of aquifers in deficit (with negative availability) for 2018 and 2020 were verified, also considering its positive counterpart, obtaining a coincidence of 81.24% for 2018 and 76.79% for 2020. If the comparison is conducted only with those aquifers with deficits, the coincidence is 55.10% and 48.72%, respectively.

## 4 Conclusions

Evaluating machine learning algorithms for predicting the available volume in aquifers and regression with random forests (RFR) obtained the best results, followed by M5', RNA, and SVMR. The advantage of RFR over M5' was expected, as the former is built using multiple trees. On the other hand, RNA and SVMR have more potential parametric combinations than RFR. In this work, the most common parametric values were used. However, learning algorithms have different sensitivity levels to their parameter values, so a deeper exploration into the tuning process (using grid search or random search, for example) could yield a fairer comparison.

The classification of aquifers predicted by RFR to fall into deficit, given the prediction of the available volume, had a coincidence of 55.10% for 2018 and 48.72% for 2020. Although the estimates were consistent, the accuracy was relatively low. It is important to note that a deficit state occurs when availability is less than 0, so quantities slightly above this threshold are not considered in deficit. Since the RFR metrics are good, slightly widening the range to determine the risk of shortfall around the numerical prediction could increase the number of matches. From this perspective, using a two-class classification technique (in deficit/with availability) could be more appropriate. Validating this approach and comparing it with the results of numerical prediction remains a future task.

From the above, it is concluded that RFR can acceptably predict water availability in aquifers in the short term but not the final deficit status. However, given the low RRSE obtained, RFR can be helpful for proactive groundwater management, improving the protection and conservation of the resource. In this sense, it is essential to note that it is not suggested that the responsibility for concessioning the extraction permits be left to a machine-learning model. The abstraction inherent in the construction process of these models can omit essential social and environmental elements, which is especially critical in water management. Relying solely on machine learning models for management poses an ethical dilemma. This process must be transparent, equitable, and responsible for all parties involved.

## References

Almuhaylan, M. R., Ghumman, A. R., Al-Salamah, I. S., Ahmad, A., Ghazaw, Y. M., Haider, H., & Shafiquzzaman, M. (2020). Evaluating the impacts of pumping on aquifer depletion in arid regions using MODFLOW, ANFIS and ANN. *Water (Switzerland)*. https://doi.org/10.3390/w12082297.

CONAGUA. (2000). *Norma Oficial Mexicana NOM-011-CONAGUA-2000 Conservación del recurso agua, Que establece las especificaciones y el método para determinar la disponibilidad media anual de las aguas nacionales.* Comisión Nacional del Agua.

CONAGUA. (2009, August 28). ACUERDO por el que se da a conocer la ubicación geográfica de 371 acuíferos del territorio nacional, se actualiza la disponibilidad media anual de agua subterránea de 282 acuíferos, y se modifica, para su mejor precisión, la descripción geográfica de 202. *Diario Oficial de La Federación de México*.

CONAGUA. (2010a, July 8). ACUERDO por el que se da a conocer el resultado de los estudios de disponibilidad media anual de las aguas subterráneas de 36 acuíferos de los Estados Unidos Mexicanos, mismos que forman parte de las regiones hidrológicas que se indican. *Diario Oficial de La Federación de México*.

CONAGUA. (2010b, July 8). ACUERDO por el que se da a conocer el resultado de los estudios de disponibilidad media anual de las aguas subterráneas de 44 acuíferos de los Estados Unidos Mexicanos, mismos que forman parte de las regiones hidrológicas que se indican. *Diario Oficial de La Federación de México*.

CONAGUA. (2010c, August 16). ACUERDO por el que se da a conocer el resultado de los estudios de disponibilidad media anual de las aguas subterráneas de 41 acuíferos de los Estados Unidos Mexicanos, mismos que forman parte de las regiones hidrológicas que se indican. *Diario Oficial de La Federación de México*.

CONAGUA. (2011a, January 25). ACUERDO por el que se da a conocer el resultado de los estudios de disponibilidad media anual de las aguas subterráneas de 50 acuíferos de los Estados Unidos Mexicanos, mismos que forman parte de las regiones hidrológicas administrativas que se indican. *Diario Oficial de La Federación de México*.

CONAGUA. (2011b, December 14). ACUERDO por el que se da a conocer el resultado de los estudios de disponibilidad media anual de las aguas subterráneas de 58 acuíferos de los Estados Unidos Mexicanos, mismos que forman parte de las regiones hidrológicas administrativas que se indican. *Diario Oficial de La Federación de México*.

CONAGUA. (2011c, December 14). ACUERDO por el que se da a conocer el resultado de los estudios de disponibilidad media anual de las aguas subterráneas de 142 acuíferos de los Estados Unidos Mexicanos, mismos que forman parte de las regiones hidrológico-administrativas que se indican. *Diario Oficial de La Federación de México*.

CONAGUA. (2013, December 20). ACUERDO por el que se actualiza la disponibilidad media anual de agua subterránea de los 653 acuíferos de los Estados Unidos Mexicanos, mismos que forman parte de las regiones hidrológico-administrativas que se indican. *Diario Oficial de La Federación de México*.

CONAGUA. (2015, April 20). ACUERDO por el que se actualiza la disponibilidad media anual de agua subterránea de los 653 acuíferos de los Estados Unidos Mexicanos, mismos que forman parte de las regiones hidrológico-administrativas que se indican. *Diario Oficial de La Federación de México*.

CONAGUA. (2018a). *Estadísticas del Agua en México 2018*. 303.

CONAGUA. (2018b, January 4). ACUERDO por el que se actualiza la disponibilidad media anual de agua subterránea de los 653 acuíferos de los Estados Unidos Mexicanos, mismos que forman parte de las Regiones Hidrológico-Administrativas que se indican. *Diario Oficial de La Federación de México*.

CONAGUA. (2020, September 17). ACUERDO por el que se actualiza la disponibilidad media anual de agua subterránea de los 653 acuíferos de los Estados Unidos Mexicanos, mismos que forman parte de las regiones hidrológico-administrativas que se indican. *Diario Oficial de La Federación de México*.

CONAGUA. (2023, November 9). *Geovisor de acuíferos*. https://sigagis.conagua.gob.mx/dma230911/

Coulibaly, P., Anctil, F., Aravena, R., & Bobée, B. (2001). Artificial neural network modeling of water table depth fluctuations. *Water Resources Research*. https://doi.org/10.1029/2000WR900368.

Crosbie, R. S., Davies, P., Harrington, N., & Lamontagne, S. (2015). Ground truthing groundwater-recharge estimates derived from remotely sensed evapotranspiration: a case in South Australia. *Hydrogeology Journal*. https://doi.org/10.1007/s10040-014-1200-7.

Daliakopoulos, I. N., Coulibaly, P., & Tsanis, I. K. (2005). Groundwater level forecasting using artificial neural networks. *Journal of Hydrology*. https://doi.org/10.1016/j.jhydrol.2004.12.001.

Fang, P., Zhang, X., Wei, P., Wang, Y., Zhang, H., Liu, F., & Zhao, J. (2020). The classification performance and mechanism of machine learning algorithms in winter wheat mapping using Sentinel-2 10 m resolution imagery. *Applied Sciences (Switzerland)*, *10*(15). https://doi.org/10.3390/app10155075.

FAO. (2011). The State of the World's Land and Water Resources for Food and Agriculture: Managing systems at risk. In *Food and Agriculture Organization of the United Nations; London: Earthscan.*

Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., & Trigg, L. (2006). Weka. In *Data Mining and Knowledge Discovery Handbook* (pp. 1305–1314). Springer-Verlag. https://doi.org/10.1007/0-387-25465-X_62.

Ganesh, N., Jain, P., Choudhury, A., Dutta, P., Kalita, K., & Barsocchi, P. (2021). Random forest regression-based machine learning model for accurate estimation of fluid flow in curved pipes. *Processes*, *9*(11). https://doi.org/10.3390/pr9112095.

Gao, L., Connor, J. D., & Dillon, P. (2014). The economics of groundwater replenishment for reliable urban water supply. *Water (Switzerland)*. https://doi.org/10.3390/w6061662.

Gonzalez-Sanchez, A., Frausto-Solis, J., & Ojeda-Bustamante, W. (2014). Predictive ability of machine learning methods for massive crop yield prediction. *Spanish Journal of Agricultural Research*, *12*(2), 313–328. https://doi.org/10.5424/sjar/2014122-4439.

Han, J., & Kamber, M. (2006). Data Mining, Southeast Asia Edition: Concepts and Techniques. *Morgan Kaufmann*.

Haykin, S. (1999). Neural networks: a comprehensive foundation second edition. In *The Knowledge Engineering Review* (Vol. 13, Issue 4).

Kanyama, Y., Ajoodha, R., Seyler, H., Makondo, N., & Tutu, H. (2020). Application of Machine Learning Techniques in Forecasting Groundwater Levels in the Grootfontein Aquifer. *2020 2nd International Multidisciplinary Information Technology and Engineering Conference, IMITEC 2020*. https://doi.org/10.1109/IMITEC50163.2020.9334142.

MacAllister, D. J. (2024). Groundwater decline is global but not universal. In *Nature* (Vol. 625, Issue 7996). https://doi.org/10.1038/d41586-024-00070-3.

Maimon, O., & Rokach, L. (2009). Introduction to Knowledge Discovery and Data Mining. In *Data Mining and Knowledge Discovery Handbook* (pp. 1–15). Springer US. https://doi.org/10.1007/978-0-387-09823-4_1.

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. https://doi.org/10.1002/widm.1301

Steyn, M. (2018). *Short-term stream flow forecasting and downstream gap infilling using machine learning techniques*.

Uc-Castillo, J. L., Marín-Celestino, A. E., Martínez-Cruz, D. A., Tuxpan-Vargas, J., & Ramos-Leal, J. A. (2023). A systematic review and meta-analysis of groundwater level forecasting with machine learning techniques: Current status and future directions. *Environmental Modelling & Software*, *168*, 105788. https://doi.org/https://doi.org/10.1016/j.envsoft.2023.105788.

UN. (2023). The United Nations World Water Development Report 2023: Partnerships and Cooperation for Water. In *Handbook of Water Purity and Quality*.

UN. (2022). United Nations World Water Development Report 2022: Groundwater, making the invisible visible. In *Significance* (Vol. 19, Issue 3). https://www.unesco.org/reports/wwdr/2022/en

Vapnik, V., Golowich, S. E., & Smola, A. (1997). Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems*.

Wang, L., Zhou, X., Zhu, X., Dong, Z., & Guo, W. (2016). Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *Crop Journal*, *4*(3). https://doi.org/10.1016/j.cj.2016.01.008.

Wang, X., Liu, T., Zheng, X., Peng, H., Xin, J., & Zhang, B. (2018). Short-term prediction of groundwater level using improved random forest regression with a combination of random features. *Applied Water Science*. https://doi.org/10.1007/s13201-018-0742-6.