

Experimental study of time series forecasting methods for groundwater level prediction

Michael Franklin Mbouopda¹, Thomas Guyet²
Nicolas Labroche³, and Abel Henriot⁴

¹ University Clermont Auvergne, Clermont Auvergne INP, CNRS, Mines
Saint-Etienne, LIMOS, Clermont-Ferrand, France

² Inria – Centre Inria de Lyon

³ Université de Tours

⁴ BRGM

Abstract. Groundwater level prediction is an applied time series forecasting task with important social impacts to optimize water management as well as preventing some natural disasters: for instance, floods or severe droughts. Machine learning methods have been reported in the literature to achieve this task, but they are only focused on the forecast of the groundwater level at a single location. A global forecasting method aims at exploiting the groundwater level time series from a wide range of locations to produce the predictions at a single place or at several places at a time. Given the recent success of global forecasting methods in prestigious competitions, it is meaningful to assess them on groundwater level prediction and see how they are compared to local methods. In this work, we created a dataset of 1026 groundwater level time series. Each time series is made of daily measurements of groundwater levels and two exogenous variables, rainfall and evapotranspiration. This dataset is made available to the communities for reproducibility and further evaluation. To identify the best configuration to effectively predict groundwater level for the complete set of time series, we compared different predictors including local and global time series forecasting methods. We assessed the impact of exogenous variables. Our result analysis shows that the best predictions are obtained by training a global method on past groundwater levels and rainfall data.

Keywords: time series · forecasting · groundwater · local and global forecasting · benchmark.

1 Introduction

Groundwaters are water bodies entrapped in soils and rocks, representing storage of significant volumes of water, usually preserved from pollutants. Generally speaking, rainfall would penetrate through soils, rocks and tends to increase the volume of groundwater, whereas drainage by springs or rivers, or water abstraction by pumping would decrease the volume of available groundwater in a given reservoir. Wells and boreholes are the devices capable of monitoring the

level of water in the reservoir, and are more generally designated under the term of *piezometers*. Forecasting groundwater levels contributes to the responsible management of an essential resource for different uses — human consumption (two-thirds of the water supplies for human consumption come from underground sources), irrigation, industrial uses (cooling, washing, etc.) — but also to watercourse flow management [21,22].

This article is focused on comparing local and global forecasting methods for predicting the evolution of groundwater levels at single or multiple locations. In France, groundwater levels records are stored in a publicly available database (ADES) operated by BRGM⁵. Data can be requested using public API's through the Hub'Eau portal <http://hubeau.eaufrance.fr/>. The database stores a network of approximately twenty thousand wells with at least one measure, but for more than a thousand of those piezometers, several years of continuous data are available, with only a few missing data. The prediction horizon was set to about three months (90 days). This corresponds to a forecast of the medium-term evolution of the groundwater levels. In terms of water management, this length of time corresponds to the needs of drought anticipation, usually performed in the late spring after the winter recharge period. In general terms, this timescale corresponds to the vast majority of the needs of applied water management.

The prediction of these levels is a true challenge given the complexity of the hydrological mechanisms at play [5]. Many numerical models have been developed for this purpose. Certain models are based on physical groundwater modelling. Physical models require their parameters to be adjusted to each situation [19]. They offer an accurate prediction solution but are generally difficult to generalize on a large scale. For more systematic prediction, the use of past groundwater level time series has been viewed for years as an essential tool for water resource planning [14]. Many machine learning models have therefore been developed to predict groundwater levels. For instance, Brédy et al. [5] put forward a modelling approach for groundwater level forecasting based on two decision-tree-based models, namely Random Forest (RF) and Extreme Gradient Boosting (XGB). Rahman et al. [21] combined wavelet transform with random forest and gradient boosting trees to predict groundwater level at scale in the city of Kumamoto in Japan. Osman et al. [12] also evaluated an XGB model but also support vector regression (SVR) and neural networks. In these studies, data includes past groundwater level values as well as rainfall and evapotranspiration data. With this information, two natural groundwater inflows and outflows are taken into account. Note that we do not have information about water usage that is difficult to collect from open resources.

Among the many state-of-the-art models, it is difficult to identify the best one for predicting the evolution of groundwater levels. Therefore, it turns out to be relevant to explore various models designed to predict the evolution of time series for the specific data provided and compare them to identify the best model.

⁵ BRGM: Bureau des Recherches Géologiques et Minières (French geological survey)

In this article, we focus specifically on the following question: “*Is it preferable to train a model on measurements from a single sensor (local model) or to train a global model to forecast the evolution of piezometric levels?*”. The usual time series forecasting approaches are designed to train models to predict future measurements from a sensor based on past measurements by the same sensor. However, a model trained on each piezometer would require a large volume of past data for each device. Such an approach would rapidly fail as the amount of available data is limited. This is also not the case for the installation of new sensors for example, and more generally it means that the model has low robustness to changes in the functioning of the hydrological system measured. The use of models trained on several piezometers can ensure greater robustness and less training effort for new piezometers. Recently, global forecasting methods (GFM) has shown astonishing results by achieving the best score on distinguished sales prediction competitions [16,17], however no prior work on groundwater level forecasting has considered these approaches yet. In aquifers, groundwater level fluctuations are not completely independent since water flows from high to low altitudes, along flow paths that depend on the aquifer geometry, porosity (the amount of connected void in the soils and rocks) and thus types of rocks, geometry of the recharge area, or discharge area, unsaturated zone thickness, etc. Inside a given aquifer, it is then expected that groundwater levels at close locations or best, along a same flow path, would have a strong co-linearity, whereas this similarity would fade with increasing distance. For distinct aquifers, it is expected that a similar groundwater level signal could be observed when input signals (rainfall, effective rainfall) are similar or close, with a certain proportion of divergence due to distinct aquifers properties. In simpler words, a proportion of redundant signal can be found along all the groundwater level records, the remaining signal being site specific. These reasons make us hypothesize that GFM would be more effective on groundwater level forecasting than local forecasting methods (LFM). Our exploration of models designed to predict the evolution of groundwater level is therefore structured around the goal of comparing two modelling strategies: local vs global methods.

For each of these types of modelling, we put forward several possible implementations and compared them with respect to the root mean squared scaled error. The rest of this paper is organized as follows: The data used in this study are described in Section 2. Section 3 presents the formalization of the studied problem and the resolution methods. The conducted experiment and the results are detailed in Section 4. This work is finally concluded in Section 6.

2 Data collection

For this study, we compiled a dataset for the period from January 2015 to January 2021 (2,221 days) for a subset of 1,026 piezometers in the French mainland. In the following, we explain the selection of the subset of piezometers. Each piezometer corresponds to a multivariate time series. The multivariate time series has three variables: the groundwater level and two exogenous weather variables:

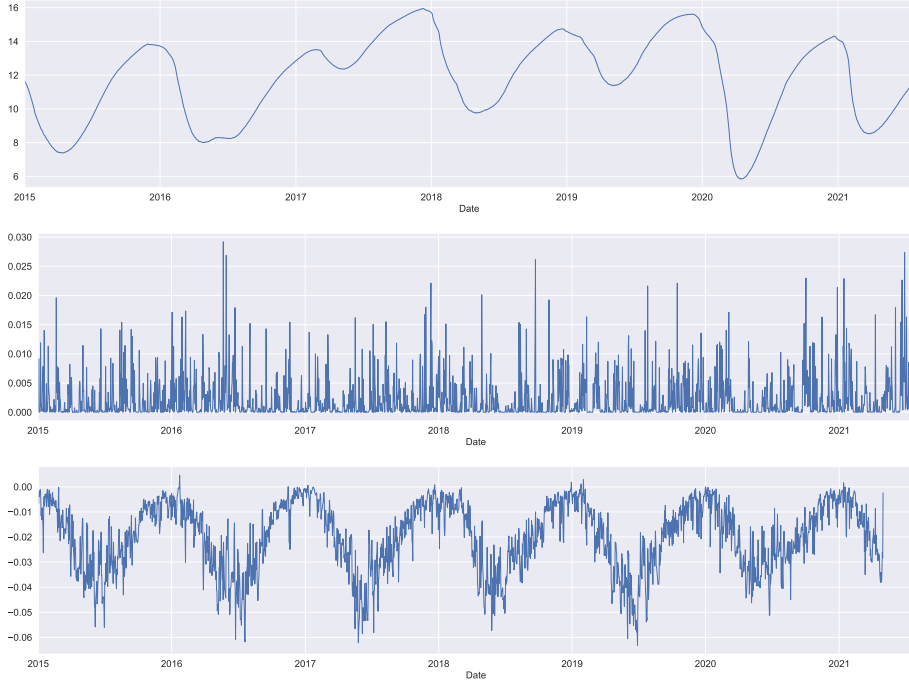


Fig. 1. From top to down, daily groundwater level, rainfall and evapotranspiration time series for the piezometer BSS000EBLL.

rainfall and evapotranspiration (ETO). These latter values have been collected from ERA5 archives using the Climate Data Store (CDS) API [10] at the location of the piezometer. This simple approach has been considered accurate enough in the context of this work, even if a more realistic approach had considered the recharge area. Nevertheless, considering a weather variable computed at the scale of each recharge area would have led to a significant increase in the complexity of the models. In addition, weather variables considered at this spatial resolution do not vary significantly with small distances, then the value at the location of the piezometer seems to be an accurate approximation. CDS provides daily weather variable measurements with a spatial resolution of 0.25° for rain and 0.1° for ETO. Hence, the three variables (groundwater level, rainfall and evapotranspiration) are available daily for each piezometer. Figure 1 shows the data for the piezometer BSS000EBLL installed in the city of *Senlis-le-Sec*. It illustrates a time series that is regular: the range of values remains between $7m$ and $16m$ all over the period, and we can identify a yearly seasonality. This is not the case for all the piezometers. Some of them do not have such seasonality and exhibit brutal changes at some dates.

In addition to groundwater level, precipitation and evapotranspiration, some past works also included sea levels, reservoir levels, and some hydrological, ge-

ological, and physiographical factors. However, we used only groundwater level, precipitation and evapotranspiration in this study as these are the three most used inputs for groundwater level forecasting [24].

The selection of the subset of piezometers has been done according to the number of missing values. The CDS data have no missing data, but piezometers may be disconnected temporarily from the network and measurements may be missed over long periods. We decided to select piezometers with less than 50 days of missing data (over the 2,221 days). This value is low and allows selecting a sufficient number of piezometers. In addition, the selected piezometers are spatially spread all over the French mainland. As a preprocessing, the missing values were attributed by linear interpolation. Thus, in the sequel, time series are considered without missing values.

Note that we did not apply any normalization of the time series values. They are provided as it to the forecasting methods.

3 Groundwater level forecasting

The groundwater levels database is a set of pairs $\mathcal{Y} = \langle Y^k, \mathbf{Z}^k \rangle$ where $Y^k : y_{1..t}^k$ is a univariate time series such that $y_i^k \in \mathbb{R}$ for any i , and $\mathbf{Z}^k : \mathbf{z}_{1..\infty}^k$ is a multivariate time series, so-called exogenous time series data ($\mathbf{Z} : \mathbf{z}_i^k \in \mathbb{R}^m$, where m is the number of exogenous time series), known as far as date t but also beyond. Time series Y^k is the groundwater level series for piezometer k , while the exogenous data series corresponds to rainfall and evapotranspiration associated to this piezometer (thus $m = 2$ in this study).

The prediction of a time series Y^k for forecast horizon h consists in estimating $y_{t+1..t+h}^k$. For this time series prediction task, the conventional data analysis approach is to construct an autoregressive model. Such a method constructs a function for the prediction of the value at date t_0 from the r latest observations of Y and \mathbf{Z} . Such a function for the prediction of the next value in the series is denoted $\varphi : \mathbb{R}^{r \times (m+1)} \mapsto \mathbb{R}$. φ can thus be seen as a regressor: it predicts a real value from the input characteristics.

Classical autoregressive models assume that φ is a linear regression (AR model), but φ can be modelled by any trainable regression function. We denote such methods as a **generalized** autoregressive model. Notice that the classical AR model is a special case of the generalized autoregressive model which assures a linear relationship between the variables.

To produce a forecast for horizon h , the prediction function is recursively applied h times. We note that it is assumed that the exogenous time series are known in the future. This problem is therefore different from that of forecasting a multivariate time series for which the prediction function outputs the value of the target series together with the values of the exogenous series. In our problem, we hypothesize that rainfall and evapotranspiration were similar from one year to the next. Consequently, forecasts can be obtained in the future by taking the mean daily values of past years. The approximation error made by

this hypothesis is preferable to that of the cumulative errors of a medium-term multivariate recursive prediction.

There exists many methods for constructing the φ function in the state-of-the-art and one goal of this work is to compare the most classical on groundwater level forecasting. The second goal is to decide whether it is better to use a local or a global method.

3.1 Local versus global time series forecasting

Given a set of time series, we denote by “local” the approaches that consider each time series as an independent dataset and build an individual model to forecast the future values of each time series. Local time series forecasting has been the de facto approach for time series forecasting. However, nowadays, many companies are collecting a large set of time series from similar sources routinely, and training a single model for each individual series is time-consuming, costly, and difficult to maintain; furthermore, possible relationships between these time series are not taken into account by local approaches. For these reasons, a new forecasting paradigm called global forecasting has emerged. A Global Forecasting Model (GFM) [13] is trained on a set of time series and is then used to forecast future values of each time series. In other words, a GFM learns to predict future values given historic values regardless of the sources of data. Global forecasting is obtaining astonishing results today, by winning prestigious competitions such as the M4 and M5 challenges [16,17] and competitions held recently on the Kaggle platform [4]. Former works on global time series forecasting claimed that time series need to be related. However, Montero-Manso and Hyndman [18] shown theoretically that whatever the heterogeneity of time series, there always exists a GFM that is as good as, or even better than any collection of local models. This result has been further supported by an experimental study [11]. Since global method has shown good results, especially in forecasting time series of sales, we are wondering if this is also the case for forecasting time series of groundwater level.

3.2 Considered methods

We explored four versions of the generalized autoregressive forecasting method (see above) as well as recent state-of-the-art neural network methods in the field. More specifically, we considered:

- Linear regression. This is the standard method for regression problems. This method is equivalent to using an AR model of order r .
- SVR [2]. Support vector regression (SVR) models address high-dimensional problems and propose non-linear models. We also used Gaussian kernels. After a few experiments, we set parameter $C = 100$.
- Random Forest [6]. This is a bagging method. We opted for a forest of 100 trees.

- Extreme Gradient Boosting [7]. This is a tree-based method which uses optimization techniques to improve the calculation efficiency of machine learning. It can thus process large datasets.
- DeepAR [23]. This is a probabilistic forecasting method based on autoregressive recurrent neural networks.
- Prophet [25]. A modular regression model with parameters that are intuitive for analysts and that can be easily adjusted using domain knowledge. Prophet is a decomposable model with three components which are trending to model non-periodic changes in the values of time series, seasonality to model periodic changes (weekly, yearly, etc.) and holidays to represent the effect of holidays on the time series values.
- NeuralProphet [26]. This is a neural network-based implementation of the Prophet model with some enhancements. In addition to the trend, seasonality and holidays components, NeuralProphet considers three additional regression components: auto-regression effect on past observations, regression effect of exogenous variables, and regression effect of lagged observations of exogenous variables. By using neural networks to learn each component, activation functions can be used to learn non-linear behaviours in the time series.

4 Experimental settings

This section presents our experimental settings. We start by presenting in detail the experiments and then give the comparison metrics.

4.1 Setup

We organized our experiments by grouping the considered forecasting models in three categories and by following a naming convention to make things easier to follow. The three categories are generalized autoregressive models, DeepAR-based models, and Prophet-based models.

Generalized autoregressive models which contains traditional models not necessarily specifically designed for forecasting: eXtreme Gradient Boosting (XGB), Linear regression (LM), Random Forest (RF), and Support Vector Machine (SVR). These regression models consider a vector as input and output a real value. Therefore, we create a dataset by sliding a window of length $r + h$ over our time series. Each position of the sliding window is a sample in our dataset with the first r values (the history) being the input and the last h (the horizon or the forecast) being the output. In the experiments, we fixed the horizon h to 93 (three months), and did not put any constraint on r , however r is generally required to be greater than h in order to expect good forecasting. We tested values from 40 to 140 and found that the best value is 100 or 110 in average – this result is in accordance with the results from [15] which suggests that a history length $1.25 \times h$ leads to the best forecasting.

DeepAR-based models composed of two models as `DeepAR` [23] is a global forecasting method. `DeepAR-L` is the `DeepAR` model trained locally, meaning that for each individual time series, a new instance of the model is trained on this time series to forecast future values of this individual time series only. On the contrary, `DeepAR-G` is a `DeepAR` model trained once on every time series to predict future values of all the time series simultaneously.

Prophet-based models made of `Prophet` [25] and `NeuralProphet` [26]. `NeuralProphet` is by default a local forecasting model, but can be configured to run as a global method. Therefore, we considered `NeuralProphet-L` (for the local version of `NeuralProphet`) and `NeuralProphet-G` (for the global version). `Prophet` can only be used as a local forecasting method.

`DeepAR`, `Prophet`, and `NeuralProphet` are specifically designed for time series forecasting: they take as input a time series and predict the user-defined number of values in the future (93 in our case). Unlike the generalized autoregressive models which use a fixed-length history, `DeepAR`, `Prophet`, and `NeuralProphet` learn from the whole past observations of the time series to make predictions.

For each forecasting method, we considered two configurations and defined a naming convention to make things easier to understand:

- the first configuration uses only historic groundwater levels in order to forecast the future levels: exogenous data are not used. This configuration is named using the forecaster name. For instance, this configuration is named `NeuralProphet-L` and `XGB` when the models used are `NeuralProphet-L` and `XGBoost` respectively.
- the second configuration does not use only historic groundwater levels to predict the future, but also historic rain and/or evapotranspiration (ETO) data as the dynamic of their corresponding phenomena could have a significant impact on groundwater level. This configuration is named as `[forecaster]+rain`, `[forecaster]+eto` and `[forecaster]+rain+eto` depending on the exogenous variables used: respectively, rain data, evapotranspiration data, or both. For instance, the configuration that uses `Random Forest (RF)` to forecast the piezometric level using historic groundwater levels and rain data is named `RF+rain`. These exogenous variables are time series that made the Z variable.

The experiment source code is written in the Python programming language using open source libraries. For `DeepAR`, we use the implementation provided in the `GluonTS` library [1]; for `Prophet` and `NeuralProphet` we use the implementation provided by the authors; we use the official implementation of the eXtreme Gradient Boosting model⁶; and for the remaining models we use `Scikit-learn` [20]. The source code is available as supplementary material and is available on a public repository⁷.

⁶ Source code and supplementary material: <https://github.com/dmlc/xgboost>

⁷ Source code: <https://github.com/frank11/piezoforecast>

4.2 Comparison metrics

The methods are compared with respect to their ability to predict the next three months of daily values. More specifically, we analysed the root mean squared error (RMSSE) of the methods over the period from the 15th of October 2021 to the 15th of January 2022, corresponding to a horizon of 93 days (roughly three months). This period is not used during the training phase of the methods. Indeed, predictions may be easier or more difficult to perform depending on the period of the year. For instance, during dry periods, groundwater is less affected by rainfall, for example. However, we focused on these three months horizon as it corresponds to the needs of drought anticipation and is also used in the literature [21]. For a time series of length n and a forecasting horizon h , the RMSSE is defined as follows:

$$RMSSE = \sqrt{\frac{\frac{1}{h} \sum_{t=n+1}^{n+h} (y_t - \hat{y}_t)^2}{\frac{1}{n-1} \sum_{t=2}^n (y_t - y_{t-1})^2}} \quad (1)$$

In addition to the RMSSE, we assess the significance of the difference between the methods. This evaluation is summarized using critical difference diagrams [8] by rejecting the null hypothesis using the Friedman test followed by a pairwise post-hoc analysis as recommended by Benavoli et al. [3]. The critical difference diagrams are drawn using the source code made public by Fawaz et al. [9]. Elsewhere, we also use box plots to summarize our results. Detailed results are available in the accompanying repository of this paper⁷.

5 Results

5.1 Generalized autoregressive models results

The performance achieved by the generalized autoregressive models are summarized on Figure 2 and the significance of the difference between each pair of models is depicted on Figure 3.

It can be observed that the best performing configurations use either only evapotranspiration as exogenous data, or no exogenous data at all. These configurations are made of linear models (see LM and LM+eto) and random forests (see RF and RF+eto). Although the difference is not significant, LM+eto performs better than LM. On the contrary, RF performs better than RF+eto. When there are no exogenous data, LM is the best model.

SVR, SVR+rain and SVR+rain+eto have very similar RMSSE in average, meaning that the impact of exogenous data is not significant for SVR-based configurations. On the contrary, exogenous data have a significant negative impact on tree-based configurations (RF and XGB) since XGB+rain and XGB+rain+eto, RF+rain, and RF+rain+eto achieve the highest RMSSE while XGB and RF are among the top four best configurations.

It is observed that the left part of Figure 3 is mainly made of configurations that use exogenous data, while the right part is mainly composed of a mix

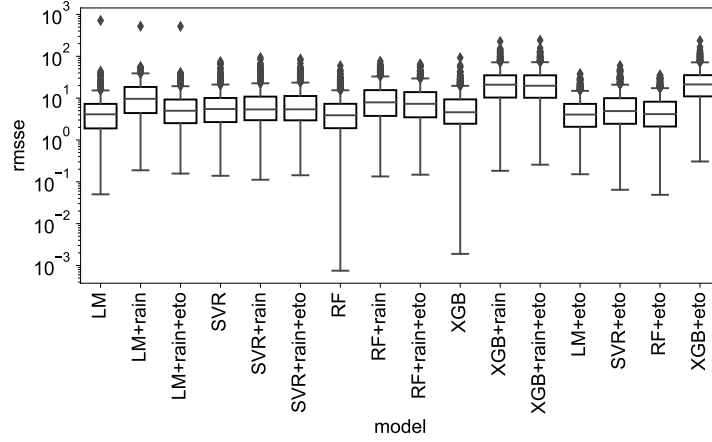


Fig. 2. RMSSE of the generalized autoregressive models.

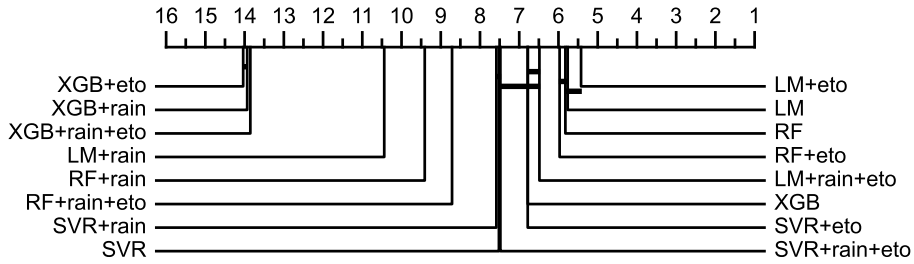


Fig. 3. Critical difference diagram of the generalized autoregressive models.

of configurations that use or do not use exogenous data. Using exogenous data decreases the performance of LM, XGB (XGB+eto, XGB+rain, and XGB+rain+eto are the worst configurations), RF, but increases performances of SVR (see SVR+eto and SVR+rain+eto). Therefore, using exogenous data makes the forecast better or worse, depending on the used forecaster and its ability to capture complex relationships.

5.2 DeepAR-based models results

The results obtained using DeepAR as local method as well as a global method are shown in Figure 4 and Figure 5.

DeepAR-L produces predictions that are significantly better than predictions of DeepAR-G, and this regardless of using exogenous data or not. This suggests that a local training strategy for DeepAR is better than a global one for ground-water level prediction.

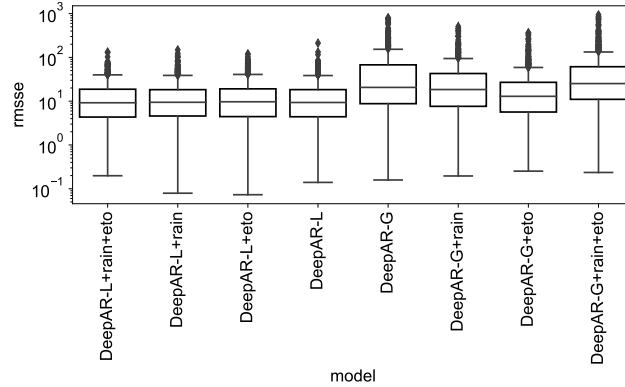


Fig. 4. Box plots summary of DeepAR models results.

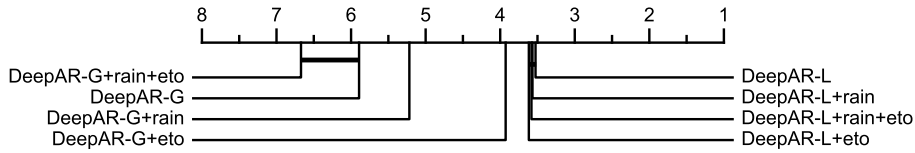


Fig. 5. Critical difference diagram of DeepAR-based models.

Although **DeepAR** is naturally designed to take advantage of exogenous data when available, it can be observed that rain and/or evapotranspiration have negative, yet insignificant impact on **DeepAR-L**'s predictions. On the contrary, exogenous data significantly improve **DeepAR-G**'s predictions, except when rain and evapotranspiration are used simultaneously.

5.3 Prophet-based models results

The comparison of **Prophet** and **NeuralProphet**-based configurations are displayed in Figure 6 and Figure 7.

Unexpectedly, **NeuralProphet-G** produces forecasts that are worse than those obtained by any other **NeuralProphet** or **Prophet**-based configurations. However, using exogenous data (rain and/or evapotranspiration) makes **NeuralProphet-G** the best forecaster, particularly **NeuralProphet-G+rain**. This result suggests that training a global **NeuralProphet** model for groundwater forecasting achieves more reliable predictions when exogenous data are used.

Unlike **DeepAR**, **NeuralProphet** works better when executed globally than locally, in particular when exogenous data are used. Furthermore, using exogenous data has a positive impact on **Prophet**, **NeuralProphet-L**, and **NeuralProphet-G**. This impact is not significant for **Prophet**, but it is for **NeuralProphet**.

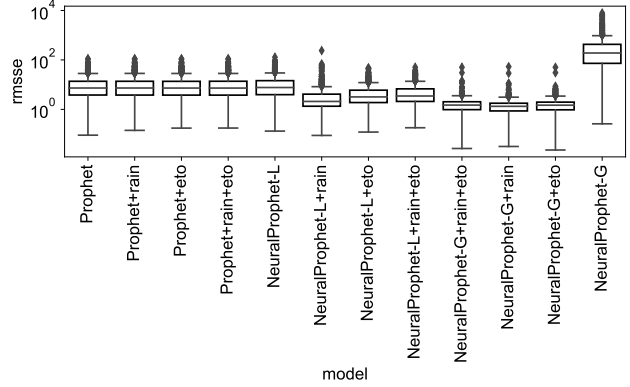


Fig. 6. Box plots summary of Prophet and NeuralProphet results.

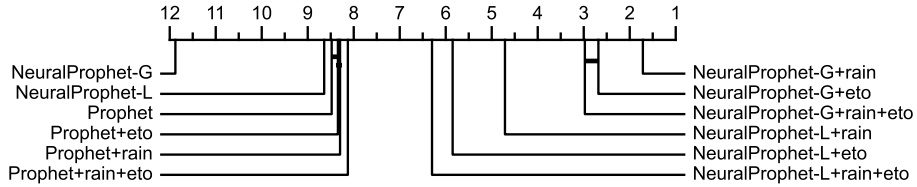


Fig. 7. Critical difference diagram of Prophet-based models.

In the absence of exogenous data, Prophet significantly outperforms NeuralProphet at predicting the future values of groundwater levels.

5.4 Comparing the three groups of models

The previous section discussed the performances of the three groups of models separately. It compares them to each other. For the sake of readability, we selected some “representative” methods from each group. However, an exhaustive comparison is available as supplementary material⁷. The representative models for each category are the two best and the two worst configurations. Also, we make sure that predictions of the selected methods are significantly different from each other regarding the critical difference diagram of the corresponding group. The comparison of the representatives methods is depicted in Figure 8.

The linear model (LM) and NeuralProphet-G respectively produce the most accurate and the least accurate forecasts when exogenous data are absent. The configuration NeuralProphet-G significantly outperforms the others when exogenous data are used, with NeuralProphet-G+rain being the best of all. The configuration NeuralProphet-G+rain is significantly better than the configuration NeuralProphet-G+eto, suggesting that precipitations impact groundwater level more than evapotranspiration does. However, as LM+eto outperforms

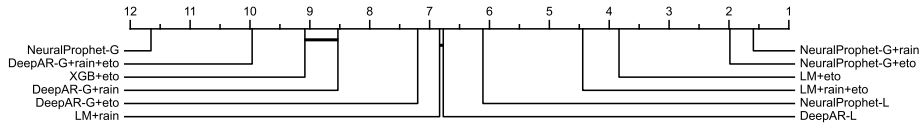


Fig. 8. Critical difference diagram comparing the representatives of groups.

LM+rain, the impact of exogenous data depends on the capabilities of the model to exploit these additional sources of knowledge.

Except for NeuralProphet-G, NeuralProphet-based configurations are better than DeepAR-based configurations at forecasting groundwater levels, especially when exogenous data are available and the training is done on every time series (i.e. globally).

DeepAR-L and LM+rain are not significantly different from each other in terms of RMSSE. The same observation is true for XGB+rain+eto and DeepAR-G+rain. This result suggests that no long-term dependency is lost by using a history of length 100. In other words, the time dependency in groundwater level data is not longer than 100 time steps when focused on the prediction at a horizon of 90 time steps.

Figure 9 shows predictions obtained with NeuralProphet-G+rain for some piezometers. We observe that forecasts are very close to ground truth groundwater levels.

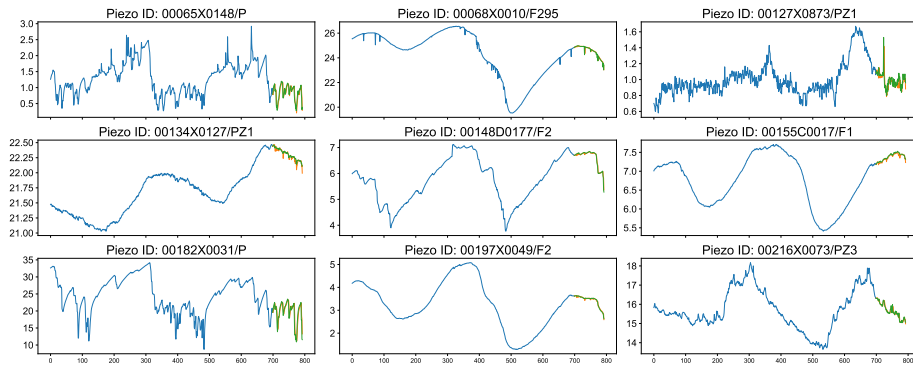


Fig. 9. Some forecasts obtained using the configuration NeuralProphet-G+rain.

5.5 Discussion

We have compared a set of state-of-the-art forecasting methods for groundwater level prediction in order to come up with the best method. We have considered

local forecasting methods as well as global methods and we evaluated the impact of using exogenous data. Our finding is that selection of the best model is highly dataset dependent. When no exogenous data are considered, LM models perform better than more complex models. Nevertheless, these models do not appear as the more useful, as they cannot efficiently afford any prediction given a set of weather prediction (drought/flood). They can still be used when no rain or evapotranspiration data are available at a given location. When considering exogenous data, `NeuralProphet-G+rain` has been found to be the best model among the set we tested. In this particular case, the global model tends to outperform local models. It catches useful information in the redundancy found in the different piezometers. Our intuition is that complex processes are at play on groundwater levels (trends, seasonality). They have to be taken into account in forecasting models, for instance with `NeuralProphet` models. In addition, this type of model is rather simple to implement and could be easily extended or enhanced for a larger dataset. Nonetheless, this training strategy does not work with all the forecaster, e.g. with `DeepAR`.

We did not consider a couple of aspects of this study and that could improve groundwater level predictions.

The first limitation is related to the choice of the methods compared and the datasets. On the one hand, we acknowledge that there exist many more forecasting methods in the literature in addition to what we considered in this study. But as we wanted to evaluate the performance of global methods trained on a large collection of groundwater level time series, we selected the most effective forecasting methods in the literature. On the other hand, the dataset used in this study is only representative of the French groundwater levels – there is no guarantee that our results generalize to a different country. Keeping this in mind, we made our source public to allow our work to be extended easily. Nevertheless, as global methods achieved the best performance, we believe that they will generalize at least as well as any local method.

The second limitation is that we did not search for the optimal hyperparameter values for each method. Instead, we used the default parameters of the implementation used. Although we could have fixed this limitation using techniques such as a grid search and Bayesian optimization. We chose a simple approach with the default values because performing hyperparameter tuning for such a large number of datasets (1,026 multiplied by the number of methods) would require a lot of time and resources. However, knowing how methods are compared to each other without hyperparameters tuning could be seen as a measure of the effort required by methods to be effective; in other words, how many are methods “plug-and-playable”.

Finally, global forecasting methods we evaluated in this work have been specifically designed for global forecasting. We have not assessed the capabilities of generalized autoregressive methods when trained globally. Although assessed methods were enough to evaluate our hypothesis about the effectiveness of GFM for groundwater level forecasting, comparing generalized autoregressive methods when trained globally with specifically designed GFM is worthy.

6 Conclusion

In this work, we performed an experimental study consisting of evaluating several methods for predicting the evolution of groundwater levels for 1,026 different piezometers. We considered local and global forecasting methods, and we compared them regarding the root mean squared scaled error metrics. The conducted experiments confirmed our hypothesis that global forecasting methods could be more effective than local ones for groundwater level prediction. In particular, global `NeuralProphet` achieved the best predictions when using past groundwater levels and rainfall as input. However, when only past groundwater levels are available (no rainfall and no evapotranspiration) local methods and particularly the linear regression model achieve the best predictions. As mentioned in the discussion section, this work could be improved. In particular, we are planning to assess the performance of generalized autoregressive methods (XGB, RF, LM, SVR, etc.) when they are trained globally and comparing them to specifically designed global methods.

References

1. Alexandrov, A., Benidis, K., Bohlke-Schneider, M., Flunkert, V., Gasthaus, J., Januschowski, T., Maddix, D.C., Rangapuram, S., Salinas, D., Schulz, J., Stella, L., Türkmen, A.C., Wang, Y.: `GluonTS`: Probabilistic and neural time series modeling in Python. *Journal of Machine Learning Research* **21**(116), 1–6 (2020)
2. Awad, M., Khanna, R.: Support vector regression. In: *Efficient learning machines*, chap. 4, pp. 67–80. APress (2015)
3. Benavoli, A., Corani, G., Mangili, F.: Should we really use post-hoc tests based on mean-ranks? *Journal of Machine Learning Research* **17**(1), 152–161 (2016)
4. Bojer, C.S., Meldgaard, J.P.: Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting* **37**, 587–603 (2021)
5. Brédy, J., Gallichand, J., Celicourt, P., Gumiere, S.J.: Water table depth forecasting in cranberry fields using two decision-tree-modeling approaches. *Agricultural Water Management* **233**, 106090 (2020)
6. Breiman, L.: Random forests. *Machine learning* **45**(1), 5–32 (2001)
7. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. pp. 785–794 (2016)
8. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research* **7**(1), 1–30 (2006)
9. Fawaz, H.I., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A.: Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery* **33**, 917–963 (2019)
10. Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., et al.: ERA5 hourly data on single levels from 1979 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS) **10** (2018)
11. Hewamalage, H., Bergmeir, C., Bandara, K.: Global models for time series forecasting: A simulation study. *Pattern Recognition* **124**, 108441 (2022)

12. Ibrahim Ahmed Osman, A., Najah Ahmed, A., Chow, M.F., Feng Huang, Y., El-Shafie, A.: Extreme gradient boosting (XGBoost) model to predict the groundwater levels in Selangor Malaysia. *Ain Shams Engineering Journal* **12**(2), 1545–1556 (2021)
13. Januschowski, T., Gasthaus, J., Wang, Y., Salinas, D., Flunkert, V., Bohlke-Schneider, M., Callot, L.: Criteria for classifying forecasting methods. *International Journal of Forecasting* **36**, 167–177 (2020)
14. Kisi, O., Shiri, J., Nikoofar, B.: Forecasting daily lake levels using artificial intelligence approaches. *Computers & Geosciences* **41**, 169–180 (2012)
15. Lara-Benítez, P., Carranza-García, M., Riquelme, J.C.: An experimental review on deep learning architectures for time series forecasting. *International Journal of Neural Systems* **31**, 2130001 (2021)
16. Makridakis, S., Spiliotis, E., Assimakopoulos, V.: The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting* **36**, 54–74 (2020)
17. Makridakis, S., Spiliotis, E., Assimakopoulos, V.: M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting* p. in press (2022)
18. Montero-Manso, P., Hyndman, R.J.: Principles and algorithms for forecasting groups of time series: Locality and globality. *International Journal of Forecasting* **37**, 1632–1653 (2021)
19. Nayak, P.C., Rao, Y.S., Sudheer, K.: Groundwater level forecasting in a shallow aquifer using artificial neural network approach. *Water resources management* **20**(1), 77–90 (2006)
20. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011)
21. Rahman, A.S., Hosono, T., Quilty, J.M., Das, J., Basak, A.: Multiscale groundwater level forecasting: coupling new machine learning approaches with wavelet transforms. *Advances in Water Resources* **141**, 103595 (2020)
22. Rodriguez-Galiano, V., Mendes, M.P., Garcia-Soldado, M.J., Chica-Olmo, M., Ribeiro, L.: Predictive modeling of groundwater nitrate pollution using random forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (southern Spain). *Science of the Total Environment* **476**, 189–206 (2014)
23. Salinas, D., Flunkert, V., Gasthaus, J., Januschowski, T.: DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting* **36**, 1181–1191 (4 2017)
24. Tao, H., Hameed, M.M., Marhoon, H.A., Zounemat-Kermani, M., Salim, H., Sungwon, K., Sulaiman, S.O., Tan, M.L., Sa’adi, Z., Mehr, A.D., et al.: Groundwater level prediction using machine learning models: A comprehensive review. *Neurocomputing* **489**(C), 271–308 (2022)
25. Taylor, S.J., Letham, B.: Forecasting at scale. *PeerJ Preprints* (2017)
26. Triebe, O., Hewamalage, H., Pilyugina, P., Laptev, N., Bergmeir, C., Rajagopal, R.: NeuralProphet: Explainable forecasting at scale. *arXiv:2111.15397v1* (2021)