Reconstruction of missing groundwater level data by using Long Short-Term Memory (LSTM) deep neural network

M.T. Vu, Abderrahim Jardani, Nicolas Massei, Matthieu Fournier

Reconstruction of missing groundwater level data by using Long Short-Term Memory (LSTM) deep neural network

M. T. Vu [1], A. Jardani [1], N. Massei [1] and M. Fournier [1]

[1] Université de Rouen, M2C, UMR 6143, CNRS, Morphodynamique Continentale et Côtière, Mont Saint Aignan, France

Abstract:

Monitoring groundwater level (GWL) over long time periods is critical in understanding the variability of groundwater resources in the present context of global changes. However, in Normandy (France) for example, GWLs have only been systematically monitored for ~20 to 50 years. This study evaluates Long Short-Term Memory (LSTM) neural network modeling to reconstruct GWLs, fill gaps and extend existing time-series. The approach is illustrated by using available monitoring fluctuations in piezometers implanted in the chalk aquifer in the Normandy region, Northern France. Here GWL data recorded over 50 years at 31 piezometers in northwestern Normandy is employed to perform GWL prediction. To optimize the network performance, the most influential factors that impact the accuracy of prediction are first determined, such as the network architecture, data quantity and quality. The resulting network is adopted to reconstruct measurements in the piezometers step by step with an increment of missing observation time. The approach requires no calibration for the time-lag in data processing and the implementation relies only on the groundwater level fluctuations to retrieve missing data in the targeted piezometers.

Keywords: Long Short-Term Memory, Groundwater level, Hydrogeology, Neural Network, Karstic aquifer, Normandy.

Introduction

Monitoring of groundwater level fluctuations over time is considered the main source of information for hydrologists to improve understanding of the evolution of aquifers and water resources across time under global change, and eventually to establish water resource management strategies. Indeed, the analysis of the groundwater level collected in a set of piezometers allows to understand the hydrodynamic behavior of the aquifer with respect to the hydrological cycle, to longer-term climate variability, anthropogenic impacts (water abstractions) and to apprehend the recharge process (Bekesi et al., 2009, Valdes et al. 2014). Interpretation of water level variations also offers the possibility of identifying preferential flow paths with high transmissivities and detecting zones, which constitute the best suited areas to be exploited, and thus protecting them against any form of contamination (Almasri and Kaluarachchi, 2007). Therefore, management plans adopted are strongly conditioned by the number of piezometers used in monitoring and their capacity both to cover the spatial heterogeneity of the aquifer and to capture the impact of climate mutations on water storage in the long and short terms. In general, these management plans are drawn up on incomplete and fragmented hydraulic databases that provide only a partial understanding of the hydrosystem due to the very high cost of installing piezometers and maintaining the acquisition of hydraulic parameters over time.

Facing the threats of climate change and the urge for establishing long-term strategies for the management and preservation of the environment, public authorities and the scientific community have been engaged in recent years in the construction of environmental databases (Anderson et al., 2008; Guo and Lin, 2016). This has led to the emergence of observatories on water resources in the world, where a large number of multi-parameter sensors are installed in the different compartments of the critical zone to monitor hydro-physical-geochemical-microbiological parameters in order to study the impact of climate change (Hipsey et al., 2015; Characklis et al, 2007; Zacharias et al, 2011; Jourde et al, 2018, Gaillardet et al., 2018). However, investigating the impact of environmental and climate

changes on hydrology requires long-term multidecadal time-series be available. This has led to a number of initiatives devoted to reconstructing long-term hydrological or hydrometeorological time-series, that are often based on the derivation of hydrological variations from long-term large-scale climate reanalyzes (e.g. Caillouet et al., 2019; Bonnet et al., 2020, Devers et al., 2020). This is also encouraged by the emergence of data science. Such research fields are driven by the progress in computer technologies, which have created a favorable environment for the development of Machine Learning techniques (Rajaee, et al., 2019). These approaches have resurfaced in recent years thanks to their ability to generate predictive models by analyzing massive data in less time. Machine learning applications in the geosciences are gaining popularity, particularly in the prediction of groundwater fluctuations, including extreme events (Mosavi et al., 2018). Indeed, deep-learning algorithms, such as neural networks, have the ability to reconstruct missing piezometric data and thus build a database that can be used as calibration data for climate projections (Sarhadi, et al, 2016). The deep learning algorithms are designed as black box models in which the input and output data are linked by a large number of weight and bias matrices defined in the neurons constituting the hidden layers (Rajaee et al., 2019). Weights and biases in deep learning models are determined in the training process through their ability to match the network response to the output of training data by using an optimization algorithm. Once the training process is complete, the validation test is performed on another subset of data not used in the training stage to check the generalization of the network, which depends on the quality and quantity of the training data as well as the architecture of the networks.

In the realm of groundwater level reconstruction, various algorithms and architectures have been tested in recent years. Lallahem et al, 2005 applied the Multilayer-Perceptron architecture (MLP) to predict groundwater fluctuations in the fractured aquifer in Northern France using precipitation, evapotranspiration and piezometric data as training data. The same approach was used by Trichakis et al. 2011 in predicting the level of the Edwards karst aquifer based on the precipitation, temperature, pumping and groundwater level data.

Ghose et al. 2018 opted for the prediction of the groundwater level with a Recursive Neural Network in which the temporal dependence of the variables is included in the prediction using the previous output at time t-1 in the prediction of the variable at time t. This incorporation of long-term

dependencies significantly improves groundwater prediction compared to the MLP network according to a comparative study conducted by Coulibaly et al. 2001. As a more advanced type of RNN, LSTM has been applied to predict the water fluctuations of the Hetao irrigation district in China by using precipitation, temperature, evaporation, and monthly water diversion recorded over 14 years (Zhang et al., 2018). By comparisons between those two networks, Bowes et al. 2019 reported a better performance of LSTM over RNN in modeling and predicting the GWL response to storm evens at Norfolk city, Virginia. LSTM was also applied for predicting the water levels of Baltic River and Long Creek with precipitation, relative humidity, mean temperature, stream level, stream flow, evapotranspiration, heat degree as input data collected over 6 years (Afzaal et al, 2019).

This manuscript discusses LSTM relevance for reconstruction of missing data from a network of 31 piezometers installed to monitor water fluctuations in a highly heterogeneous porous, fractured and karstic regional aquifer of Upper Normandy (France). These piezometric fluctuations describe a strong spatio-temporal heterogeneity of water dynamics in the aquifer over 50 years. Here, reconstruction is based exclusively on the use of GWLs collected over long periods on the same piezometers to fill the short (2 years) and long (47 years) gaps in the other piezometers.

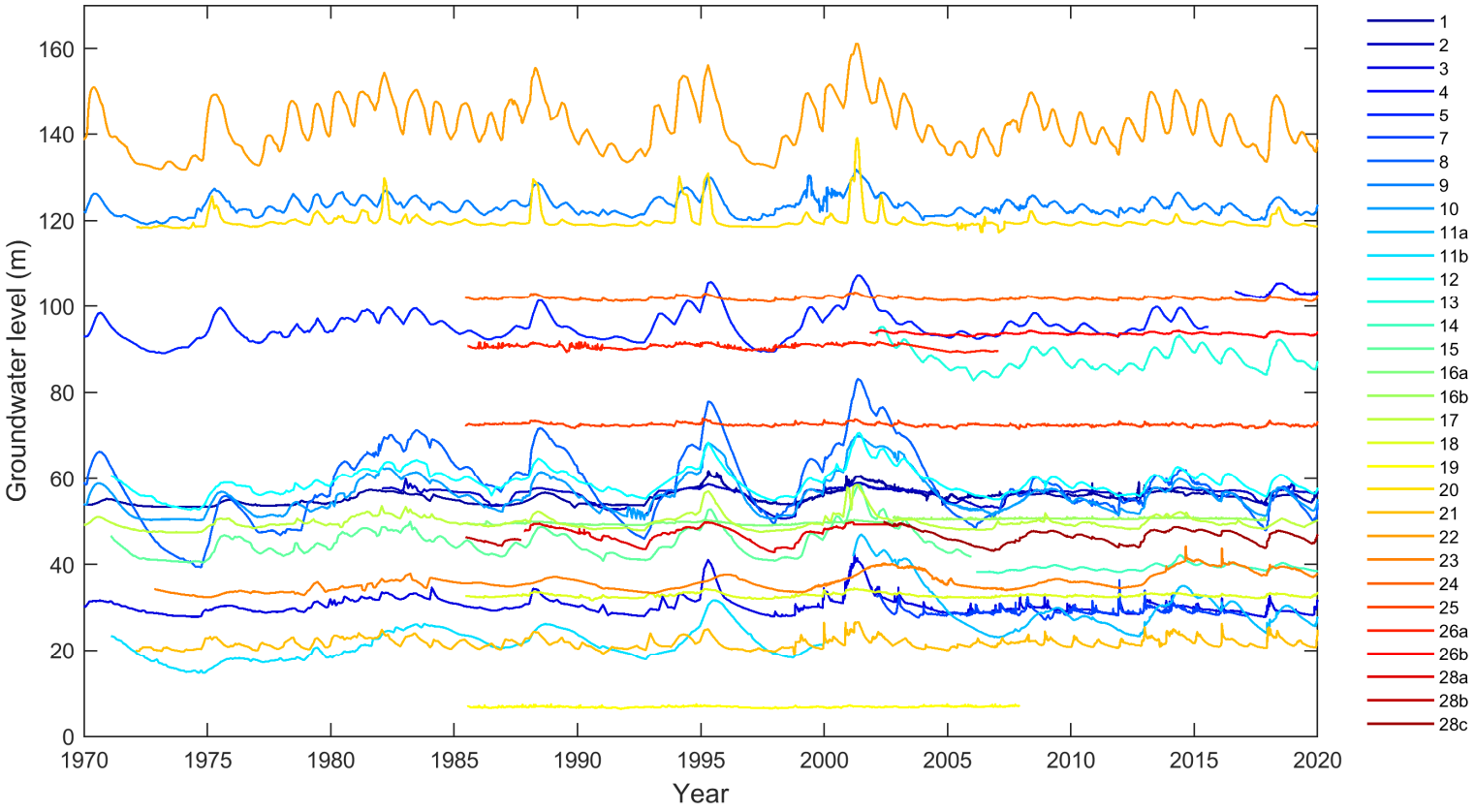
2. Study area and Data acquisition

The efficiency of the LSTM algorithm in the historical reconstruction of groundwater fluctuation data will be studied from its application on a set of hydraulic data acquired on a part of the piezometric network used in the monitoring of the Normandy karstic basin. This basin is highly karstified and constitutes the main source of water supply in the region (Slimani et al, 2009). The study area is located in northwestern Normandy (France), where 31 piezometers were installed at different dates to monitor piezometric fluctuations in the Normandy karstic aquifer (**Figure 1**). This aquifer is unconfined and is covered by superficial geological layers of loess and clay with flints that are disturbed by the presence of sinkholes and crypto-sinkholes (Jardani et al., 2006; Valdes et al,

2014). This geological cover controls the recharge processes of the karstic aquifer with rapid infiltration through sinkholes and slow infiltration through the superficial formation that forms a perched aquifer during rainy periods. These two dynamics can be observed in the records of groundwater level on piezometers and water flux on karst springs (El Janyani et al, 2013; Valdes et al, 2014).

Figure 1: Investigated zone on the left-wing of Seine River, Normandy, France. The white points represent 31 GWL observation piezometers.

The hydraulic data analyzed in this study are composed of 31 time series of water levels recorded on a set of piezometers distributed over the area (**Figure 2**). However, the duration of the records is not uniform, some records are long and last almost 50 years, as in the case of 8 piezometers, and others are very short and cover only 3 to 5 years. In addition, some piezometers have been withdrawn from monitoring networks in recent years (i.e. piezometers N°5, N°11b, N°15), but they provide valuable information on past groundwater fluctuations. **Figure 3** provides details on the duration of each series. The sampling frequency is also not uniform, with weekly measurements for the first 35 years (1970-2005) and daily measurements for the last 15 years.



117 **Figure 2: Monitoring groundwater level in 31 piezometers in 50 years.**

118

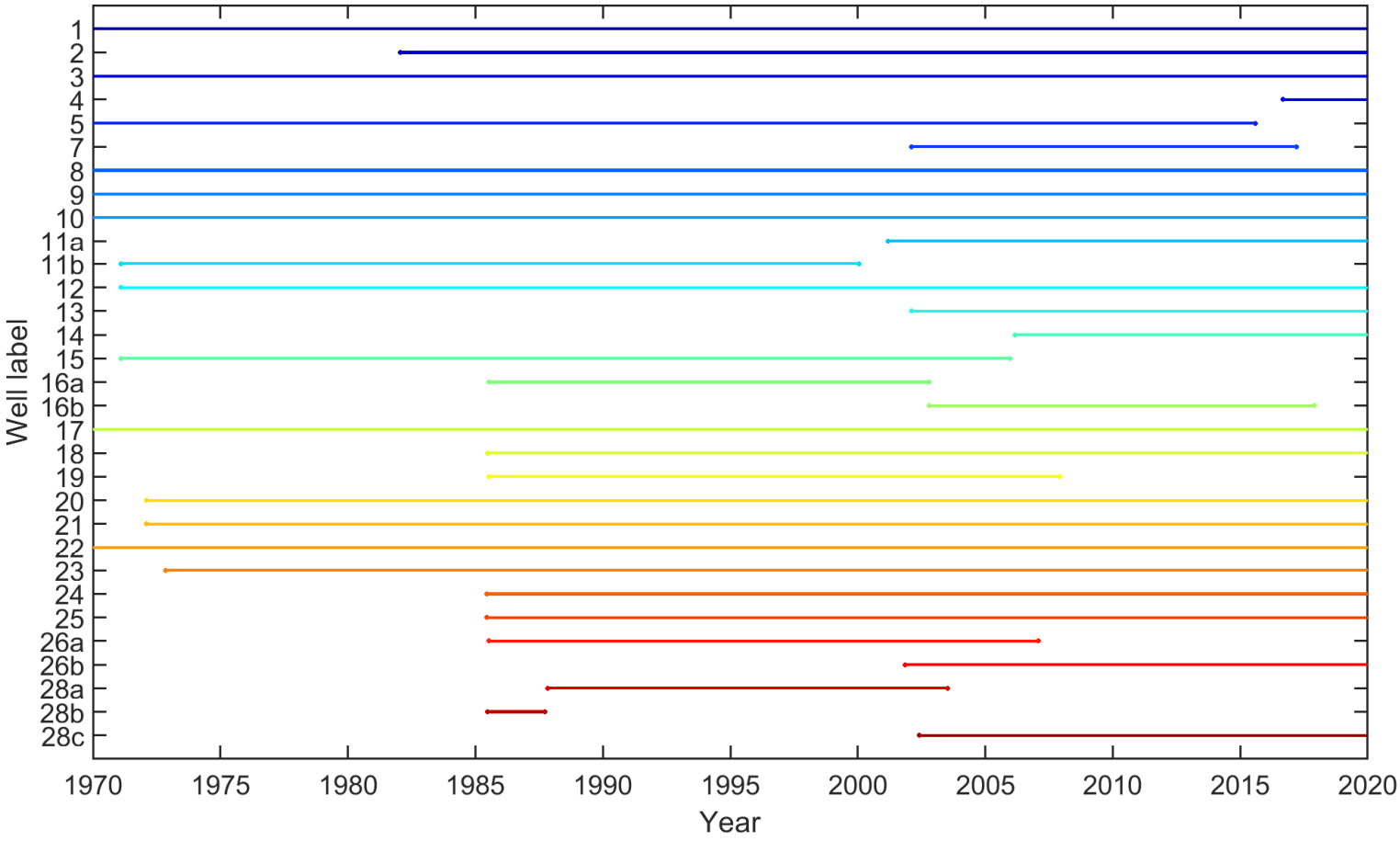


Figure 3: Measurement time-window at 31 piezometers over 50 years from 1970 to 2020 (the same legend with Figure 2).

The mean of GWL fluctuations observed in these piezometers can reach to 30 times of difference which proves the existence of a strong hydraulic gradient particularly between the upstream and downstream compartments. The amplitude of the fluctuations over time also varies significantly, from about 0.1 to 10 m. These contrasts in the fluctuations are due to the degree of karstification of the hydrosystem. Hydraulic measurements share and carry certain common characteristics linking to the regional climate conditions, which can allow the deep learning algorithm to establish links between the GWL data measured in various piezometers without requiring the incorporation of the precipitation signal in this process.

3. Methodology and Model design

3.1 Long Short Term Memory

Long short term memory (LSTM) is an enhanced architecture of the Recurrent Neural Network (RNN) that has been designed to process the time-dependent variables presented in time series (Hochreiter & Schmidhuber, 1997). This type of network has the advantage of incorporating past information into the prediction of the future state of the variable when the input data have certain dependencies. In the RNN, the memory effect is taken into account by using an unrolled loop cell which allows the previous information to flow into the prediction of the next step. However, the way it is structured does not allow an effective processing of long-term dependencies as its learning process leads to the vanishing gradients during the back-propagation. To overcome this obstacle, the LSTM networks have been developed with an efficient structure comprising three gates: input gates, output gates, and forget gates that ensures the preservation of previous information with a stable gradient calculation (see **Figure 4**). On these three gates within a cell state, the information is processed by a sequential computation using the following equations (Hochreiter & Schmidhuber, 1997; Felix et al, 2000):

$$\begin{cases} i_t = \boldsymbol{\sigma}(W_i x_t + U_i h_{t-1} + b_i) \\ f_t = \boldsymbol{\sigma}(W_f x_t + U_f h_{t-1} + b_f) \\ o_t = \boldsymbol{\sigma}(W_o x_t + U_o h_{t-1} + b_o) \\ C_t = tanh(W_c x_t + U_c h_{t-1} + b_c) \\ C_t = f_t \otimes C_{t-1} + i_t \otimes C_t \\ h_t = o_t \otimes tanh(C_t) \end{cases}, \quad (1)$$

where x_t designates the input variable at the current time step, h_t is the output of the previous cell, C_{t-1} is the previous cell state which provides the past information. These parameters are used with a set of the weight matrices and bias vectors in the logistic sigmoid $\boldsymbol{\sigma}$, and *tanh* functions at the input, forget and outputs gates. All these weights and bias vectors are estimated during the learning process in matching the training data by using ADAM optimizer.

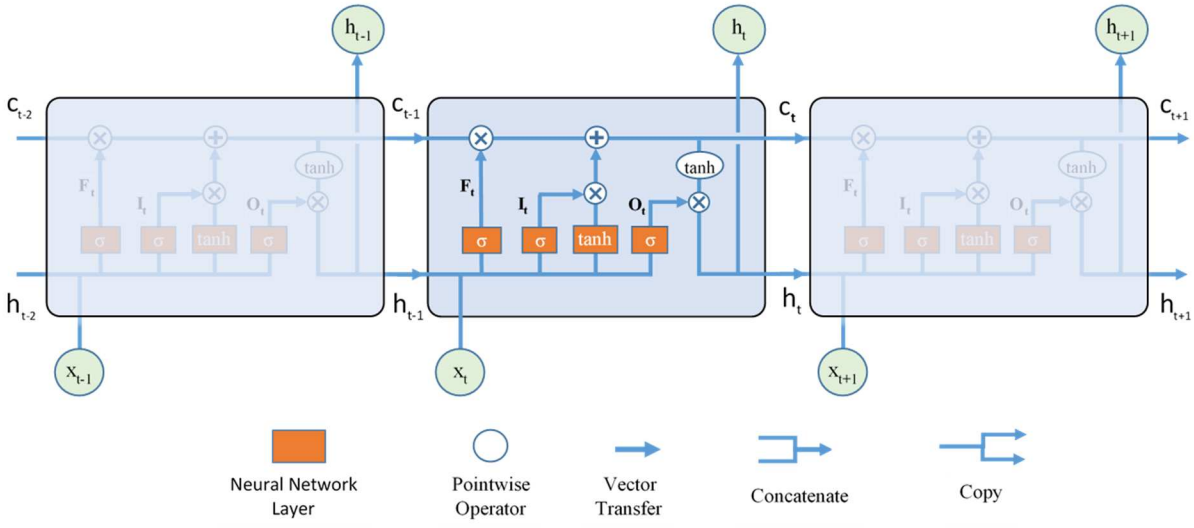


Figure 4: LSTM architecture with F, I, O denote the three gates as forget, input and output gates, respectively; x, h and c correspond to the input, output and update state of each cell (node), respectively (Sagheer & Kotb, 2019). A single hidden layer is adopted, which consists of n_{hid} nodes.

Regarding the choice of the optimal structure of LSTM networks to obtain accurate predictions, a single hidden layer is adopted in which number of nodes is determined according to the following rule (modified from Lallahem et al, 2005):

$$(n_{in} + 1) \times n_{hid} + (n_{hid} + 1) \times n_{out} \leq \frac{1}{\alpha} n_{train} \,, \tag{2}$$

where n_{in} denotes the number of nodes in the input layer, n_{out} is the number of nodes in output layer and n_{hid} is number of nodes in the hidden layer, n_{train} is number of training data, α is a coefficient, varies from 1 to over 10. In this study, to avoid overfitting, α is given a value higher than 2 as the training data doubles the degrees of freedom in the training process. In this study, the training applies ADAM algorithm with constant learning rate of 0.002 and a longest sequence option in the mini-batch. The learning takes a half of minute in Matlab running in a Dell Precision T5810 with a single GPU NVIDIA Quadro K2200.

3.2 Data calibration

Before the start of the training process, the dataset of groundwater level in the 31 piezometers from 1970 to 2020 are resampled to a weekly time step and normalized with following formulation:

$$\tilde{h}_i^m = \frac{h_i^m - \bar{h}^m}{\max_{i=1,n} |h_i^m - \bar{h}^m|} \quad , \tag{3}$$

where h_i^m, $\bar{h}^m$ are the measurement i and the mean of the whole series of n measurements in the piezometer m, respectively. This normalization facilitates learning by re-scaling all data between plus and minus one as plotted in **Figure 5**.

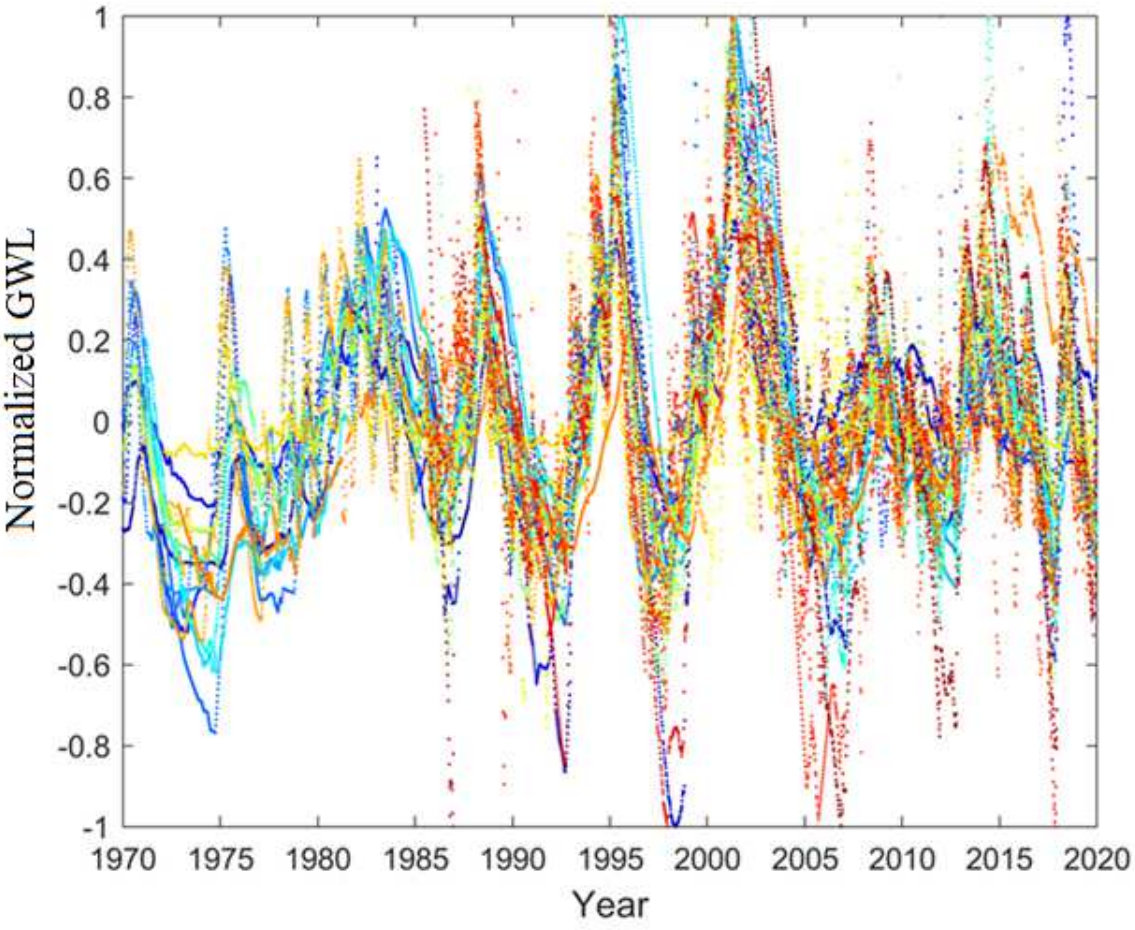


Figure 5: Normalized Groundwater level in 31 piezometers over 50 years from 1970 to 2020. The normalized fluctuations clearly share a similarity in low frequency features (the same legend with Figure 2).

Even the original time series show a changing range of amplitudes; their normalized fluctuations show similar low-frequency behaviors. These similarities can be exploited to learn how to reconstruct missing piezometric data or extrapolate data. Indeed, to achieve a reliable prediction, the training data must contain sufficient information and characteristics that are representative of the variability of the hydraulic data in the piezometer to be predicted. For this reason, the choice of the data portion that will be used to feed the networks is a first step in the process, in terms of data availability and their concordance with the targeted piezometer data.

Hence, in order to examine the impact of training data on the quality of predictions, different strategies will be discussed where various piezometers will be integrated into the predictions.

The quality of these predictions is assessed and analyzed using three criteria:

The root-mean-square error

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{h}_i - h_i)^2}{n}}, \quad (4)$$

where $\hat{h}_i$ and h_i are predicted and observation groundwater levels, n is number of testing values.

The correlation coefficient:

$$Corr_{Test} = \frac{\sum_{i=1}^{n}(\hat{h}_i - \bar{\hat{h}})(h_i - \bar{h})}{\sqrt{\sum_{i=1}^{n}(\hat{h}_i - \bar{\hat{h}})^2}\sqrt{\sum_{i=1}^{n}(h_i - \bar{h})^2}}, \quad (5)$$

where $\bar{\hat{h}}$ and $\bar{h}$ are average predicted and observation groundwater levels of testing data.

In some tests, the observation is referred in a neighboring piezometer to validate the prediction in the targeted piezometer and the correlation between two series is defined as

$$Corr_{m-k} = \frac{\sum_{i=1}^{n}(h_i^m - \bar{h}^m)(h_i^k - \bar{h}^k)}{\sqrt{\sum_{i=1}^{n}(h_i^m - \bar{h}^m)^2}\sqrt{\sum_{i=1}^{n}(h_i^k - \bar{h}^k)^2}}, \quad (6)$$

where h_i^m, h_i^k are measurement i in the piezometer m and piezometer k and their means are $\bar{h}^m$ and $\bar{h}^k$.

3.3 Validation of the Network:

The objective of this study is to use a statistical learning approach to effectively predict groundwater level fluctuation in a piezometer by analyzing hydraulic data acquired from a network of piezometers used in the learning process. This network of piezometers consists of 11 piezometers distributed over the study area with the longest observation time scale, as shown in **Figure 6**. This dataset is representative of the main hydrodynamic characteristics observed in all the piezometers studied, with mean water levels varying from 20 to 150 m and amplitudes ranging from 0.2 to 10 m.

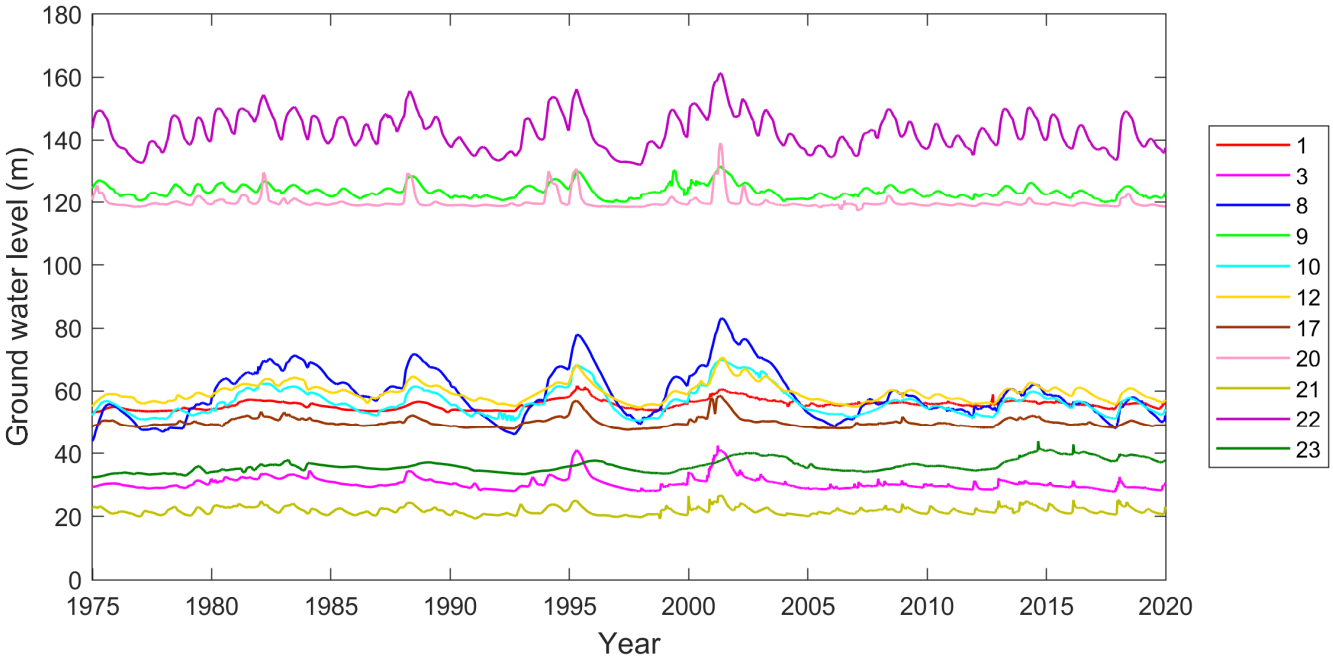


Figure 6: Observation GWL in the subset of 11 piezometers over 45 years from 1975 to 2020.

Graphical analysis of these time series shows a certain similarity in the fluctuations, particularly during major events characterized by a significant rise or fall in water level (such as the maxima of years (1976, 1988, 1996 and 2002) and the other minima of years (1977, 1992 and 2006). This proves that the piezometers are located in the same hydrological compartment, although the amplitude and response time to these extreme events remain different and are strongly influenced by the local hydrological contexts, such as the degree of karstification around the piezometers, and the hydraulic

connectivity of the piezometers with the surface karstic sinkholes. In general, the signatures of low-frequency fluctuations are shared on the majority of piezometers, while high-frequency fluctuations show a behavior that is difficult and complex to identify over time. For more details, the analysis of correlations between data series over time is presented in **Figure 7**. The analysis is carried out over the first and last 15 years in order to clarify the relationship between piezometers in space and time.

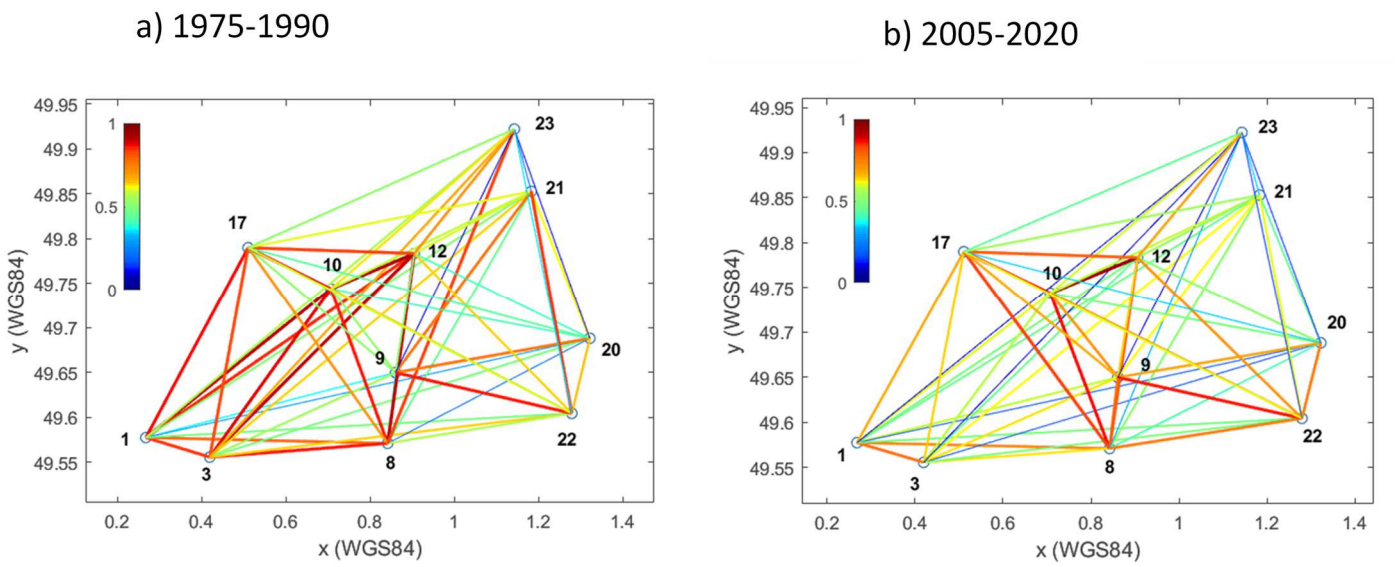


Figure 7: Correlation between groundwater level data observed at the piezometers in 1975-1990 and 2005-2020. Red/thick lines correspond to the highest correlation equal to 1, Blue/thin lines represent a low correlation equal or under 0 (according to Eq.6). Correlations between the fluctuations in the studied piezometers are spatially heterogeneous and significantly change in time.

The correlations established over the first 15 years of data (1975-1990) show the presence of a strong relationship between the data from neighboring piezometers located in the western part of the study area: piezometer 1-3-8-9-10-12-17, with a correlation coefficient varying between 0.8 and 1. However, the spatial proximity between piezometers does not always imply the presence of a correlation, as shown by the case of the 20-21 piezometers, which have a poor correlation despite their proximity.

In general, the presence or lack of correlation is an element primarily associated with the hydrodynamic conditions of the environment in which the piezometers are installed, which in turn are

controlled by hydraulic conductivity. In this context where porosity conditions the dynamics of the flows, two transfer modalities can take place: the first one is fast and is focused on the karstic conduits which are more frequently connected with the sinkholes and the second one is slow and diffuse in the matrix and small cracks. Therefore, piezometers placed on karstic conduits are characterized by a high temporal variability, carrying information on short-term rainfall events in addition to the long-term climatic variations that may also appear in piezometric records with low transmissivity. The climatic information that is the origin of these fluctuations can be expressed on all the piezometers but in a different way in terms of amplitude and time response, therefore the use of the correlation criterion to select the piezometers to be served in learning is not a trustworthy criterion. As these correlations between piezometers have been altered over time due to the accentuation of the anthropogenic factors related to urbanization, agricultural practice and intensive exploitation of water resources. These factors can lead to significant changes in time and space in the recharge process and regional groundwater flows.

All these considerations lead us to not rely on correlation as a selection criterion for training piezometers and to believe that all piezometers can be used in this process. In anyway, the impact of the initial correlation between the input and output data on the quality of the predictions will be explored as well as the impact of the number of hidden layer, on the prediction case of the GWL in piezometer N°10 using neighboring piezometers.

a) Impact of correlation between input and output data:

To understand the influence of the correlation between the input and output data on the accuracy of the prediction, the observation in a single piezometer is used to predict its neighbor, as shown in **Figure 8**. The input data come from the observations in piezometer N°9, N°3 and N°12 successively. The training data are the 15-year (1975-1990) observations, with a correlation between input and output in these three piezometers of 0.54, 0.89 and 0.94, respectively. The corresponding testing data are longer which composes observations in 30 years (1990-2020). LSTM networks adopts a single

hidden layer which consists of 25-65 nodes (α = 10) and their results are detailed in **Table 1**. The evaluation criteria for these predictions show their dependence on the initial quality of the correlation between input and output, with a clear improvement of the prediction when the data are well correlated (**Figure 8a, b, c**). However when the correlation is relatively low, the prediction only identifies the overall fluctuation trend of the target piezometer. Therefore, it is reasonable to use a piezometer with correlated data if the purpose of the prediction is to retrieve the details of the high-frequency fluctuations. The results also highlight the influence of the changing correlation over time. In **Figure 8b**, the prediction is badly reconstructed for the last 15 years 2005-2020, while the correlation between input and output (N° 3 and N°10) during this test period is much weaker than for training (1975-1990), as is the prediction for the case (N°1 and N°10) as shown in **Table 1**. This can be explained by the fact that not all the features of aquifer dynamics in recent decades have been captured in the model derived from earlier years' data. On the other hand, the correlation (N°12 and N°10) remains almost identical over the two periods (see **Figure 7**), so that the learning process is complete, which explains the good prediction of the fluctuations in **Figure 8c**.

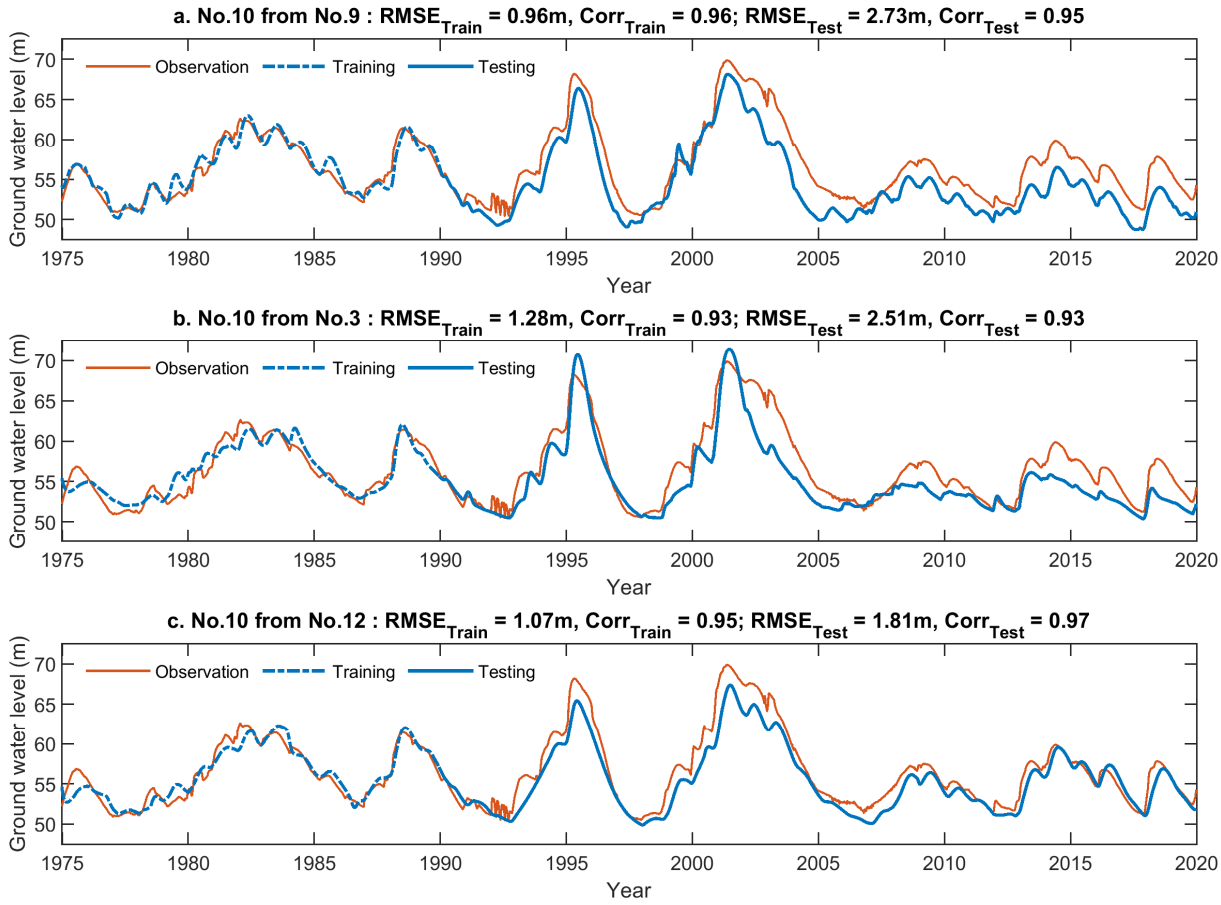


Figure 8: Predictions of GWL fluctuation in piezometer No.10 from observation data in a single piezometer. The training data is for 15 years, from 1975 to 1990, to predict 30 years from 1990 to 2020 (testing). Results are shown with increment of correlation between the targeted piezometer (output) and the feeding variable (input), the input are from: a. Piezometer N°9, b. Piezometer N°3, c. Piezometer N°12. The accuracy is directly determined by the correlation between the input and output data over time. When the correlation is low, the result reproduces only the tendency, higher correlation data result in better predictions. Predictions from single series is of limited accuracy.

b) Impact of amount of training data

In this section, the effect of the number of piezometers used in the training process is examined in relationship with the reliability of predictions. In this case, the learning will include more GWL data from neighboring piezometers to predict the water fluctuations in the piezometer N°10. Thus, three predictions are performed corresponding to three distinct training data sizes consisting of 3, 6 and 10 piezometers. The prediction results from these multiple time series are reported in the **Figure 9** and the **Table 1** show a clear improvement in the accuracy over those obtained from a single piezometer. This tendency is identifiable in **Table 1** for following tests: 1st 2nd 3th 4th – 7th 8th 9th – 12th 13th – 15th.

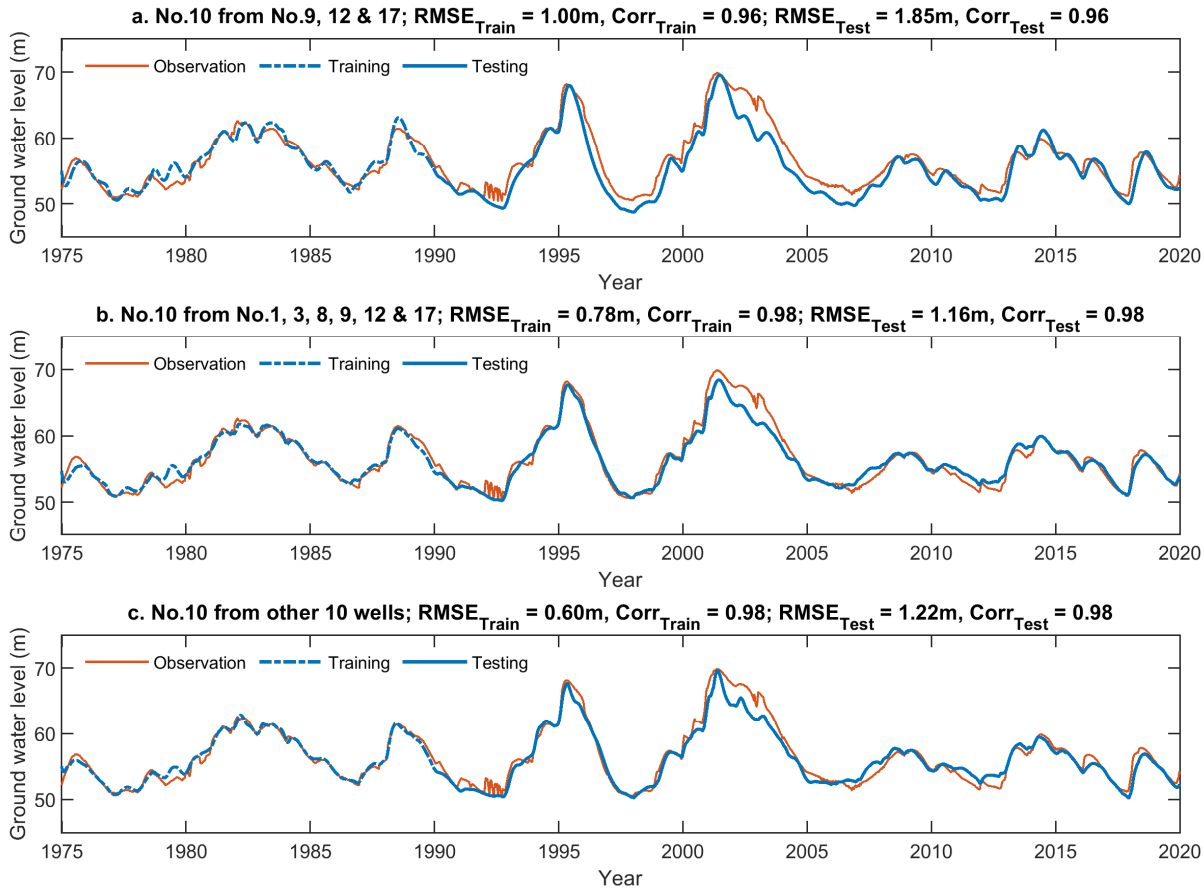


Figure 9: Prediction GWL in piezometer N°10 from number of observation series in neighboring piezometers. More input variables seem to provide better prediction in the targeted piezometer when more information is considered in the learning process, but it may also import noises in the predictions.

The use of multiple series provides a good generalization because they offer the possibility to learn from multiple features that could not be seen with a single piezometer. However, this proportionality

between the increase in the number of series involved in learning and the improvement in the quality of the predictions is not always respected. As shown in **Table 1**, particularly for predictions involving 6 piezometers in Test 16, which are not as good as those of Test 12 obtained only using 3 piezometers. The conclusion to be drawn from these analyses is that the quality of the prediction is not only conditioned by the volume of training data, in some cases the incorporation of unrelated data can negatively affect the accuracy because it will disturb the appropriate information. Therefore, data quality is also a crucial factor in prediction that should not be underestimated. In this case the combination of 10 piezometers in the training provides an accurate prediction ($Corr_{Test}$ = 0.98) of the water level in piezometer N°10, so this strategy will be adopted to predict the other piezometers dealt in the following sections.

Table 1: Comparison of RMSE (in m) for various data feeding and network configurations to predict GWL at the piezometer N°10.

Test	*Data feeding from piezometers N°*										*Network*			
	1	3	8	9	12	17	20	21	22	23	α = 3	α = 5	α = 10	α = 15
1	x										4.15	4.13	**4.09**	**4.06**
2		x									2.75	2.67	**2.51**	**2.63**
3				x							2.69	**2.48**	2.73	3.36
4					x						1.92	1.93	**1.81**	1.83
5								x			3.34	3.37	**3.29**	3.50
6									x		2.64	**2.56**	2.66	2.80
7	x	x									1.96	**1.79**	2.13	2.37
8			x	x							2.39	2.14	**2.09**	**2.06**
9					x	x					2.04	**1.86**	**1.85**	**1.84**
10							x	x			**3.35**	3.42	3.58	3.46
11									x	x	2.51	**2.20**	2.47	2.57
12	x	x	x								1.71	1.66	**1.64**	**1.60**
13				x	x	x					1.95	**1.92**	1.99	**1.90**
14							x	x	x		2.89	**2.84**	3.09	3.07
15	x	x	x	x	x	x					1.14	**1.08**	**1.12**	1.25
16					x	x	x	x	x	x	2.10	**1.96**	**1.94**	1.97
17	x	x	x	x	x	x	x	x	x	x	1.52	**1.28**	**1.26**	1.39

310 ***c) Impact of number of nodes in the hidden layer***

311 In this section, the effect of the number of nodes constituting the hidden layers on the quality of
312 predictions is analyzed. In general, the choice of neural network properties is often made through
313 trial-and-error analysis. For that the performance of networks is tested using several numbers of nodes
314 by changing the coefficient α from 3 to 15 (see Eq.2) and the results are listed in **Table 1**. The number
315 of nodes in the hidden layer is dependent on degrees of freedom in the training process. When the
316 node in the layer is relatively small, this implies a low degree of freedoms that may prevent the
317 networks to reach a convergence. However, in the opposite case with an important degree of freedoms
318 can lead to overfitting issue. As this investigation is done with a number of input variables is 3 to 15
319 times higher than the degrees of freedom, the accuracy of the predictions varies only slightly with α,
320 with an optimal value around 5-10.

321 # 4. Result and Discussion

322 After the sensitivity analysis of the LSTM networks and the implementation of the best strategy for
323 predicting groundwater levels, this section is devoted to the application of this strategy to the recovery
324 of missing hydraulic data over 50 years in the piezometric network of the study area. In this network,
325 some observations presenting a short discontinuity will be first completed from the available data and
326 later all exploited to predict shorter observations in other piezometers.

327 ***4.1. Test 1: Complete series with a short discontinuity***

328 Some series with minor gaps of 2 to 5 years is firstly filled, and then these completed series are
329 used to predict other series with significant observation gaps. These predictions are based on learning
330 complete data from seven piezometers. Some representative prediction results are shown in **Figure**
331 **10**, and the test evaluations are detailed in **Table 2**.

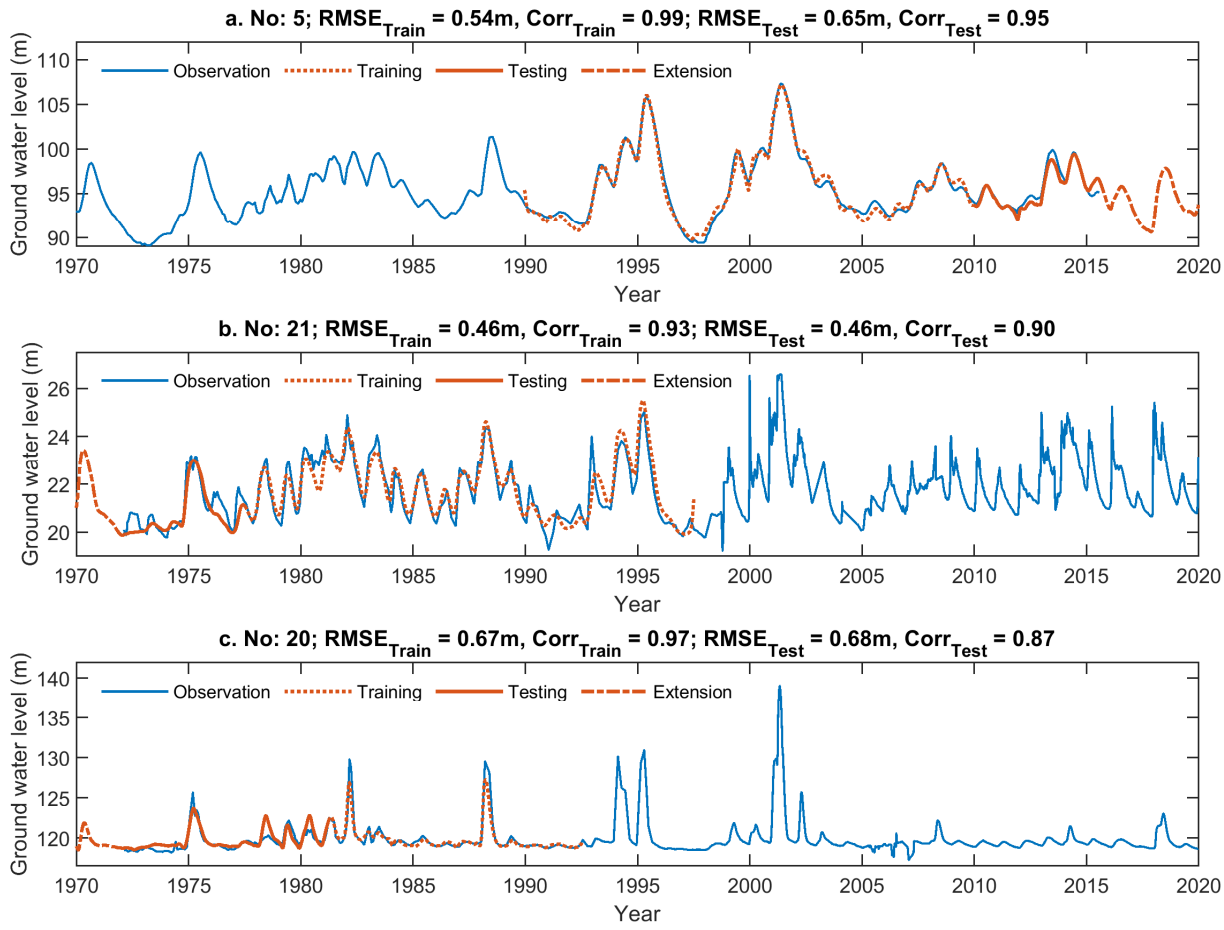


Figure 10: Complete the observations with a short missing of 2-4 years. Results are for piezometers: a. Piezometer N^o 5, b. Piezometer N^o 21 and c. Piezometer N^o 20. Predictions employ LSTM network, feeding from full-range observations in 7 piezometers (including piezometer N° 5 to optimize the feeding data). The network composes a hidden layer of 80-290 nodes (α = 3-10). The predictions can generally track the complexity of patterns in the piezometers.

Table 2: Details of the calibrations in 7 piezometers.

Target piezometer	*Data feeding from piezometers N°*								*Network*			*Result*	
	1	3	5	8	9	10	17	22	Training	Testing	n_{hid}	$RMSE_{Test}$ (m)	$Corr_{Test}$
No.5	x	x		x	x	x	x	x	1990-2010	2010-2016	80	0.65	0.95
No.12	x	x	x	x	x	x	x	x	1980-2000	1971-1980	275	0.73	0.96
No.20	x	x	x	x	x	x	x	x	1981-1992	1972-1981	155	0.68	0.87
No.21	x	x	x	x	x	x	x	x	1977-1997	1972-1977	275	0.46	0.90
No.23	x	x	x	x	x	x	x	x	1979-2000	1973-1979	290	0.54	0.86

The LSTM network is trained using all the hydraulic data acquired over 15-20 years, then tested to complete the observations over 6-8 years before extending the prediction to complete the sequence. The network includes a hidden layer with 80-290 nodes (α about 3-10). To optimize use of data, the prediction is first made for piezometer N°5 from 7 available piezometers (N°1, 3, 8, 9, 10, 17, 22), the predictions of N°12, 20, 21, 23 will then include data of N°5 in the training (the training data do not covered the predicted data of N°5, see **Table 2** for details). The predictions obtained are globally of satisfactory accuracy when the trained models arrive to identify the complexity of the hydrodynamic features in the predicted piezometers. The recovered time series are relatively short compared to the total length of the series; however, this reconstruction is crucial to provide more data to predict other series with significant gaps. In the next step, the data from these 12 piezometers are recalled extracting shorter observations from other piezometers.

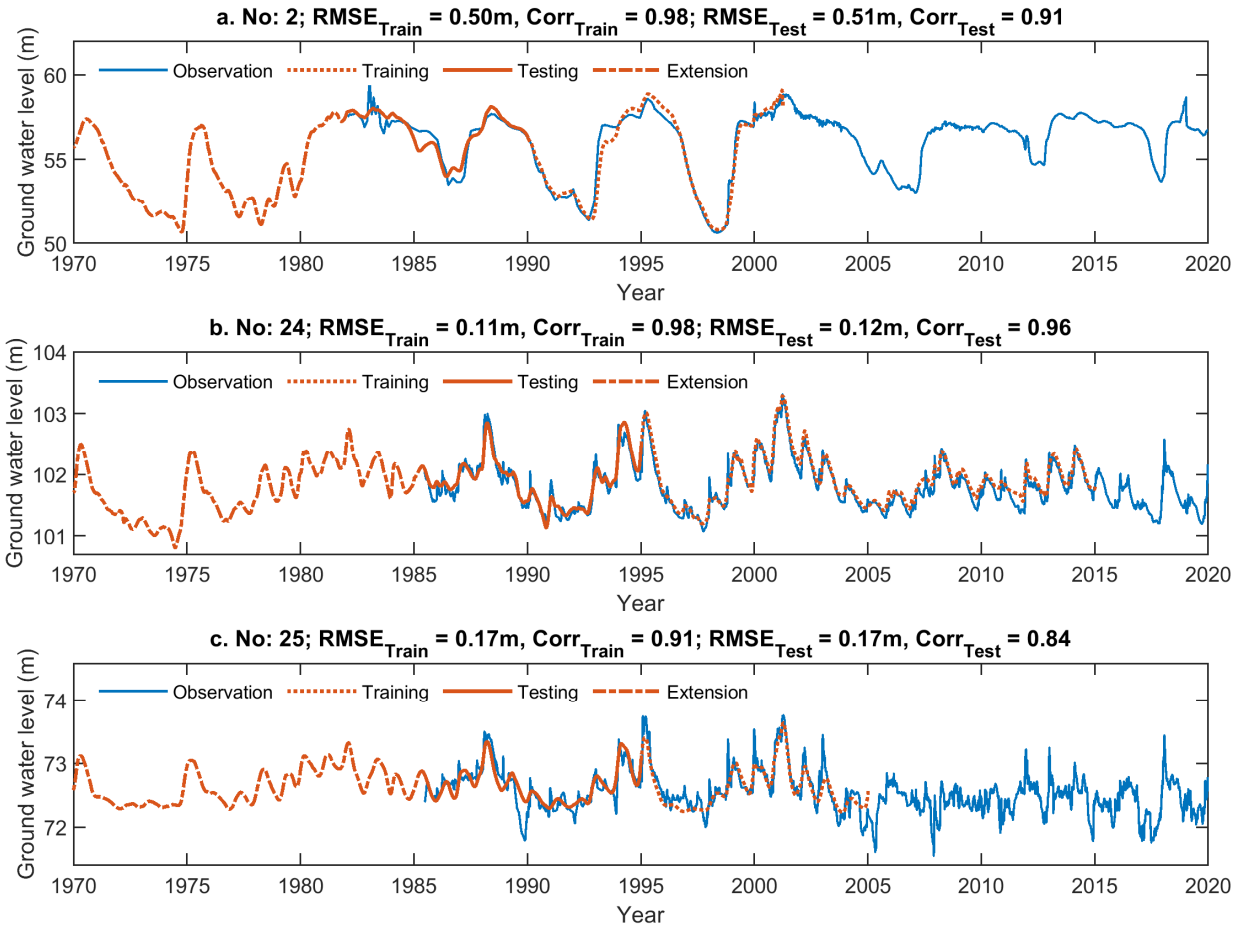


Figure 11: Complete the observations with a missing of 15 years. Results are for piezometers: a. Piezometer Nº 2, b. Piezometer Nº 24 and c. Piezometer Nº 25. Predictions employ LSTM network (α = 3-10), feeding from dataset in 12 piezometers. Accuracy of predictions do not expose to depend on the location of the piezometers when the learning process only bases on the relationship among feeding sequences.

364

365 **Table 3: Details of calibration to predict GWL from dataset in 12 piezometers.**

Target piezometer	*Network*			*Result*	
	Training	Testing	n_{hid}	$RMSE_{Test}$ (m)	$Corr_{Test}$
No.2	1990-2002	1982-1990	50	0.51	0.91
No.18	1995-2010	1985-1995	130	0.20	0.89
No.24	1995-2015	1985-1995	295	0.12	0.96
No.25	1995-2005	1985-1995	40	0.17	0.84

366

367 In this test, missing measurements in 4 piezometers using 12 piezometers are predicted by applying
368 the previous configuration. The LSTM network is trained by data of 10-20 years and its efficiency is
369 tested over 8-10 years (see **Figure 11**), details of the quality of the predictions on the test data are
370 presented in **Table 3**. The predictions are highly correlated with observed fluctuations with small
371 errors (by RMSE). Despite the fact that the series studied have different characteristics and changing
372 also over time, the training data contains a large amount of information to decrypt the complexities of
373 each type of data and to provide a satisfactory reconstruction of missing data.

374 ***4.2. Test 2: Predict long missing periods***

375 In this section, the series of eight piezometers are completed with a long void in records of nearly
376 30 years using data from 16 piezometers that include both real and recovered data. The network is
377 trained from the 10-year data and validated on the 10 years of data. In **Figure 12**, some representative
378 time series for the 8 predicted piezometers is illustrated, and details of the prediction quality of the
379 data tests are shown in **Table 4**. As noted above, neighboring piezometers are often well correlated
380 and share the same hydrogeological characteristics. In this test, these neighboring piezometers is not
381 included in the training, but use them for validating the prediction as shown in **Figure 12b&c**. The
382 predictions obtained are reliable according to the evaluation criteria and their comparisons with
383 neighboring piezometers.

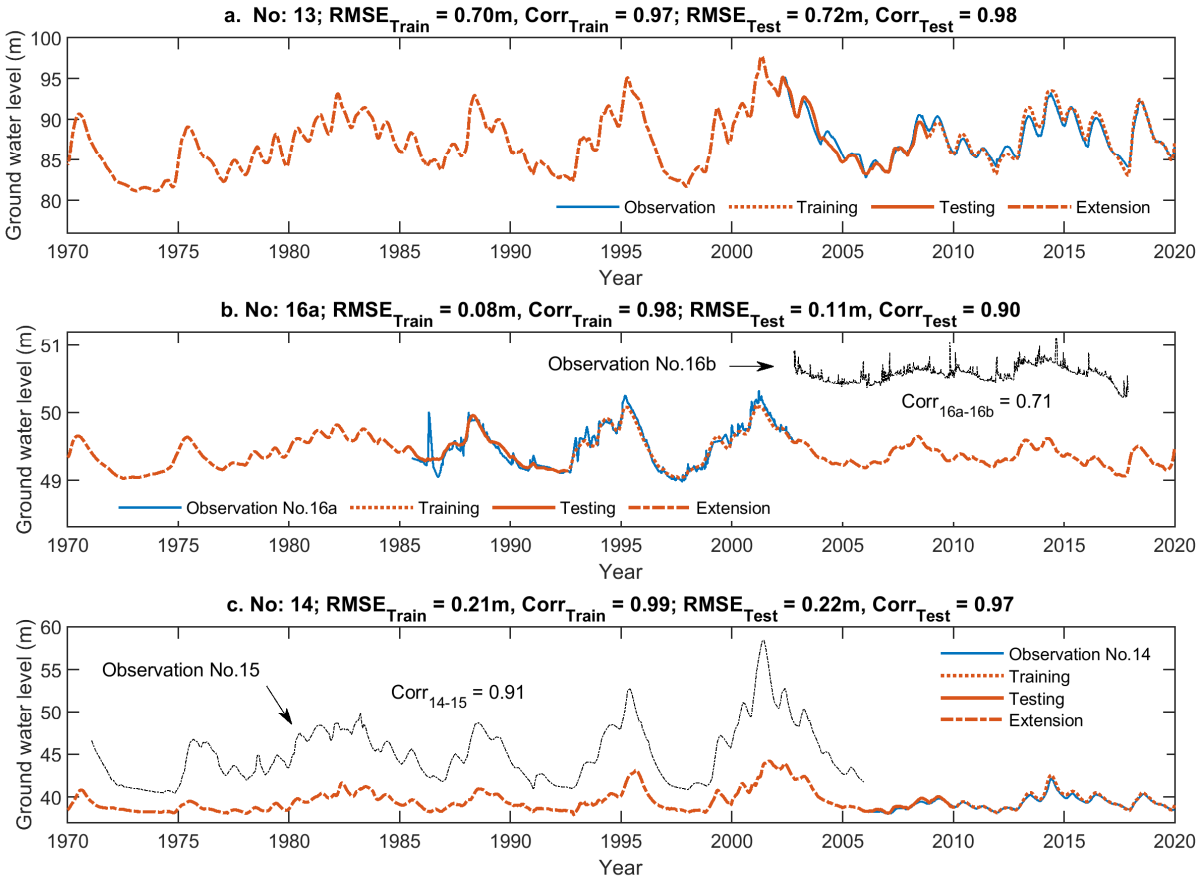


Figure 12: Complete the observations with a long missing of about 30 years. Results are for piezometers: a. Piezometer N° 13, b. Piezometer N° 16a and c. Piezometer N° 14. Predictions employ LSTM network (α = 3-10), feeding from dataset in 16 piezometers. The resulted predictions are accurate in term of RMSE, correlation coefficient and comparative to the neighbors.

Table 4: Details of calibration to predict GWL in 8 piezometers from 16 dataset.

Target piezometer	*Network*			*Result*	
	Training	Testing	n_{hid}	$RMSE_{Test}$ (m)	$Corr_{Test}$
No.16a	1993-2003	1985-1993	45	0.11	0.90
No.16b	2007-2017	2002-2007	75	0.07	0.75
No.11a	2010-2020	2002-2010	75	1.32	0.99
No.11b	1980-2000	1972-1980	150	1.22	0.90
No.13	2010-2020	2002-2010	85	0.72	0.98
No.14	2010-2020	2007-2010	75	0.22	0.97
No.7	2010-2017	2002-2010	115	0.75	0.74
No.15	1990-2005	1972-1990	135	1.08	0.91

The prediction result from piezometer 16a reveals a good reconstitution of the observed data, but it is also affected by a sudden and unexplained increase in the water level in 1987. This local effect in the test data did not occur in the training data. It is therefore a limitation of the predictive model that can be remedied by using additional metrological data such as rainfall and temperature that are not available in this study. The comparison with its neighbor shows a concordance in low frequency fluctuations. On the other hand, it is difficult to establish a correlation between the high-frequency fluctuations that may be intrinsic characteristics of each piezometer. For this reason, the correlation coefficient between them is not relatively high ($Corr_{16a\text{-}16b} = 0.71$).

All of these results will be used to predict the fluctuations in the remaining 7 piezometers in the next section.

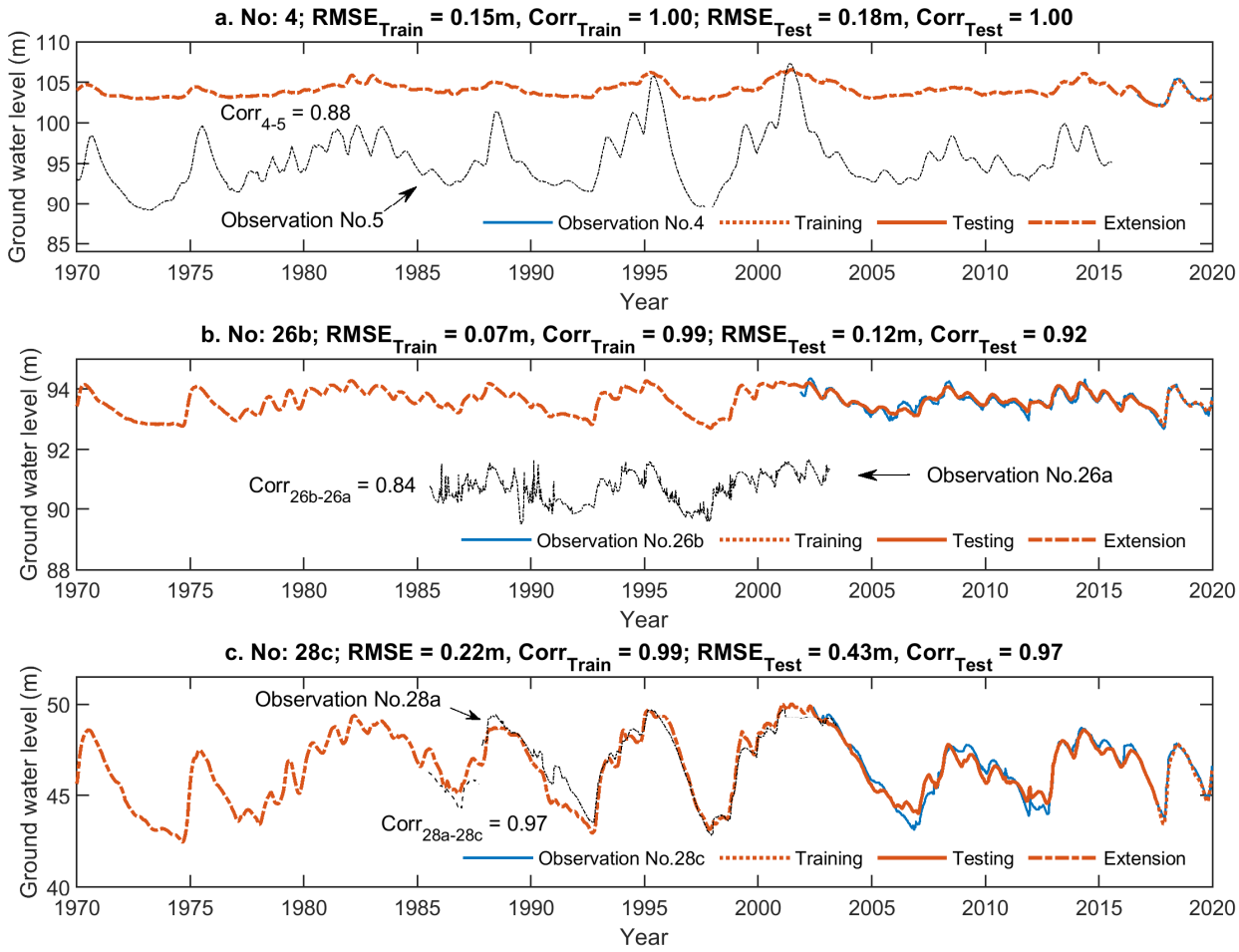
4.3. Test 3: Long predictions from very short observation for training

Previous reconstructions have allowed us to build a hydraulic database with 24 piezometers that are sufficient to attempt the recovery of the abandoned piezometers where recordings were made over a very short period of 3 years (2017-2020). In this sort of prediction with few data is so difficult to establish a generalization and validation of the networks. To avoid this issue, the lack of data in the

target piezometer will be counterbalanced by the use of observation in nearby piezometers. The validation of the approach is conducted on long series with only a short period of data used in the learning process. Some results of this type of prediction on time series with few data are shown in **Figure 13** and calibration details are shown in **Table 5**.

The discussion begins with the predictions of piezometer N° 4 where only three years of data are available and used in training and the validation is done by a comparison with data from neighboring piezometer N° 5, which is excluded from training data. The result is shown in **Figure 13a**, where the reproduction of training data is excellent, but this does not mean that the generalization is also perfect. To verify the quality of these predictions, some of them are compared with the data from the nearest piezometer. This comparison shows a correlation coefficient $Corr_{4\text{-}5} = 0.88$, which means that the prediction is successful. The data from piezometer N° 5 is not directly exploited in the training of this test, but it has been used in previous recoveries for other piezometers, so some of these features may be included indirectly in this training operation.

The same strategy is adopted for the remaining piezometers N° 26b and N° 28c, which are located at different places in the study area. The network is established using 2.5 years of data, but in this case, the available data covers 17 years, allowing us to retain sufficient data for validation. Data from a nearby piezometer is also incorporated into the validation process and the results are shown in **Figure 13b&c**. The prediction is encouraging even using only short data in the formation, the network was able to match the test period fluctuations with a correlation of 0.92 and 0.97 respectively. The prediction is also consistent with neighboring piezometers with Corr = 0.84 and 0.97, respectively. The tests prove that this approach can be applied to make long predictions from very short training data.



437

438 **Figure 13: Long predictions of 50 years from very short training of 3 years data. Results are**

439 **for piezometers: a. Piezometer N^o 4, b. Piezometer N^o 26a and c. Piezometer N^o 28c. Predictions**

440 **employ LSTM network (α = 2-3), feeding from dataset in 24 piezometers. The approach enables**

441 **to perform long predictions with a favorable accuracy from a very short training dataset.**

442

Table 5: Details of calibration to predict GWL in 7 piezometers from 24 dataset.

Target piezometer	*Network*			*Result*	
	Training	Testing	n_{hid}	$RMSE_{Test}$ (m)	$Corr_{Test}$
No.4	2017-2020	2016-2017	55	0.18	1.00
No.26b	2017-2020	2002-2017	55	0.12	0.92
No.28c	2017-2020	2002-2017	55	0.43	0.97
No.26a	1992-2005	1985-1992	285	0.27	0.78
No.28a	1993-2003	1985-1993	240	0.55	0.97
No.28b	1986-1987	1985-1986	35	0.13	0.95
No.19	1992-2005	1985-1992	200	0.11	0.64

This section is closed by discussing the reconstruction of the data in piezometer N°19, which represents a low correlation on the validation data (Corr = 0.64). As mentioned in section 4.1, using the combined hydraulic data from several piezometers in the learning process increases the risk of including impertinent data that will be a source of noise. Despite this, the prediction for No. 19 is made with a Corr = 0.64 which is slightly higher than the maximum correlation of the observed data with respect to the others (0.57).

The measurements recorded on this piezometer N° 19 are poorly correlated with the observations from the other piezometers (average correlation 0.41) may be related to the fact that piezometer N°19 is located in a hydrogeological context that is different from the rest of the piezometers used in the training. This can be explained by its proximity to the sea, which has a high tidal range of up to 8 m. In fact, this piezometer has the lowest piezometric mean compared to the rest of the 31 piezometers. In comparison with the neighboring piezometer N°17 which is closer to the sea (3 km from the coast), but its average of 49.7 m is seven times higher than that of piezometer N°19. It can therefore be concluded that the piezometer N°19 has a different hydrodynamic behavior from the piezometers involved in the training and to reconstruct its fluctuations it is necessary to re-form a network with piezometers located on the downstream part at the interface with the sea. This driving influence of

tidal regimes on GWL at coastal regions are also confirmed in other studies by Bowes et al. 2019 at Virginia US and Taormina et al. 2012 at Venice lagoon Italy.

Conclusion

In this paper, the approach adopts the LSTM which is one of the most efficient deep learning algorithms in time series processing to identify missing data in groundwater records. The approach was applied to a set of hydraulic data collected with part of the piezometer network installed to monitor groundwater fluctuations in the karstic aquifer in Upper Normandy. These piezometers have different recording durations: some have a long recording duration of 50 years, others have been abandoned so they only have a very short recording duration of 3 to 10 years. The implementation of the LSTM relies on the use of water level data in some piezometers to form and build the network to retrieve missing data from other piezometers. This tool does not require the use of meteorological data such as rainfall and temperature in training operation. The main points that emerge from the use of LSTM in this study are as follows:

- ✓ This approach is relevant for reconstructing the GWL fluctuation with satisfactory accuracy over long periods of time from even very short observations with a correlation coefficient varied from 0.64 to 0.99 and RMSE from 0.07 m to 1.08 m.
- ✓ The accuracy of the predictions depends on the quality of the training data, such as the initial correlation between input and output, as well as the duration and number of piezometric series used in the training. It is therefore crucial that the piezometers are in the same hydrogeological context and that they share certain hydrodynamic characteristics to facilitate reconstruction. The use of unreliable data in the learning process with contrasting characteristics what are observed on the predicted piezometers will lead to a poor prediction.

484 ✓ The properties of the network in terms of number of layers and number of neurons can
485 also influence the predictions and the choice of these parameters can be obtained by
486 analyzing the quality of the predictions of different configurations.

487 The hydraulic head outcome of the model can be employed to enhance predictions of GWL itself in
488 the future, associate the modelling of other processes in the subsurface, such as flow field or predict
489 the transport of contaminant in the groundwater. In future works, the prediction can be extended to
490 calibrate for every point in considering a hybrid model where a mathematic model probably joints
491 with a physic-based approach including information of surface/subsurface conditions in the real field.
492 Finally, considering the scarcity of long-term observational groundwater data and the complexity of
493 generating simulations using classically used (physics-based or conceptual) modeling approaches, it
494 will also be critical to explore the capabilities of deep learning techniques for long-term
495 reconstruction of groundwater levels. This is mandatory to tackle the issue of understanding the
496 impact of low-frequency climate variability and climate change on water resources availability.

497

References:

Afzaal H., Farooque A., Abbas F., Acharya B. and Esau T., 2020. Groundwater Estimation from Major Physical Hydrology Components Using Artificial Neural Networks and Deep Learning. *Water*, 12, 5.

Almasri M.N.; Kaluarachchi J.J., 2007. Modeling nitrate contamination of groundwater in agricultural watersheds, *Journal of Hydrology*, 343, 3-4, 211-229.

Anderson S.P., Bales R.C., Duffy C.J., 2008. Critical Zone Observatories: Building a network to advance interdisciplinary study of Earth surface processes, *Mineralogical Magazine*, 72, 1, 7-10.

Bekesi G., McGuire M., Moiler D, 2009. Groundwater Allocation Using a Groundwater Level Response Management Method - Gnangara Groundwater System, Western Australia. *Water Resource. Manage,* 23, 1665–1683.

Bonnet R., Boé J., Habets F., 2020. Influence of multidecadal variability on high and low flows: the case of the Seine basin. *Hydrology and Earth System Sciences*, 24, 1611-1631.

Bowes, B.D., Sadler, J.M., Morsy, M.M., Behl, M., Goodall, J.L., 2019. Forecasting groundwater table in a flood prone coastal city with Long Short-term Memory and Recurrent Neural Networks. *Water*, 11, 1098.

Caillouet, L., Vidal, J.-P., Sauquet, E., Graff, B., Soubeyroux, J.-M., 2019. SCOPE Climate: a 142-year daily high-resolution ensemble meteorological reconstruction dataset over France. *Earth System Science Data*, 11, 241-260.

Characklis G.W., Reed P.M., Minsker B.S., 2007. The role of the systems community in the National Science Foundation's environmental observatories, *Journal of Water Resources Planning and Management*, 133, 1.

Coulibaly, P., Anctil, F., Aravena, R., Bobee, B., 2001. Artificial neural network modelling of water table depth fluctuations. *Water Resour. Res.,* 37 (4), 885–896.

Devers, A., Vidal, J.-P., Lauvernet, C., Graff, B., Vannier, O., 2020. A framework for high-resolution meteorological surface reanalysis through offline data assimilation in an ensemble of downscaled reconstructions. Quarterly Journal of the Royal Meteorological Society.

El Janyani S., Dupont J-P, Massei, N., Slimani S., Dörfliger N., 2013. Hydrological role of karst in the Chalk aquifer of Upper Normandy, France. *Hydrogeology Journal*, 22, 3, 663–677.

Felix A.G., Jürgen S., Fred C., 2000. Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, 12, 10, 2451-2471.

Gaillardet J., Braud I., Hankard F., Anquetin S., Bour O., Dorfliger N., De Dreuzy J.-R., Galle S., Galy C., Gogo S., 2018. OZCAR: The French network of critical zone observatories, *Vadose Zone Journal*, 17, 1, 1-24.

Ghose D., Das U., Roy P., 2018. Modeling response of run off and evapotranspiration for predicting water table depth in arid region using dynamic recurrent neural network. *Groundwater for Sustainable Development,* 6, 263–269.

Guo L., Lin H., 2016. Critical zone research and observatories: Current status and future perspectives. *Vadose Zone Journal*, 15, 9.

Hipsey M.R.; Hamilton D.P., Hanson P.C., Carey C.C., Coletti J.Z., Read J.S., Ibelings B.W., Valesini F.J., Brookes J.D., 2015. Predicting the resilience and recovery of aquatic systems: A framework for model evolution within environmental observatories. *Water Resources Research*, 51, 9, 7023-7043.

Hochreiter Sepp, Schmidhuber Jürgen, 1997. Long short-term memory. Neural Computation. 9 (8): 1735-1780.

Jardani A., Dupont J.P., Revil A., 2006. Self-potential signals associated with preferential groundwater flow pathways in sinkholes. *Journal Geophysical Research*, 111, B09204.

Jourde H., Batiot-Guilhe C., Bailly-Comte V., Bicalho C., Blanc M., Borrell V., Bouvier C., Boyer J.-F., Brunet P., Cousteau M., 2011. The MEDYCYSS observatory, a multi scale observatory of flood dynamics and hydrodynamics in karst (Mediterranean border Southern France). *Advances in the Research of Aquatic Environment*, 551-560.

Lallahem S., Mania J., Hani A., Najjar Y., 2005. On the use of neural networks to evaluate groundwater levels in fractured media. *Journal of Hydrology*, 307, 1-4, 92-111.

551 Mosavi A., Ozturk P., Chau K., 2018. Flood prediction using machine learning models: Literature
552 review. *Water*, 10, 11, 1536.
553 Rajaee T., Ebrahimi H., Nourani V., 2019. A review of the artificial intelligence methods in
554 groundwater level modelling. *Journal of Hydrology*, 572, 336-351.
555 Sagheer A., Kotb M., 2019. Time series forecasting of petroleum production using deep LSTM
556 recurrent networks, *Neurocomputing*, 323, 203-213.
557 Sarhadi A., Burn D.H., Johnson F., Mehrotra R., Sharma A., 2016. Water resources climate change
558 projections using supervised nonlinear and multivariate soft computing techniques. *Journal of*
559 *Hydrology*, 536, 119-132.
560 Slimani S., Massei N., Mesquita J. et al., 2009. Combined climatic and geological forcings on the
561 spatio-temporal variability of piezometric levels in the chalk aquifer of Upper Normandy (France) at
562 pluridecennal scale. *Hydrogeology Journal*, 17, 1823.
563 Taormina, R., Chau, K.-W., Sethi, R., 2012. Artificial neural network simulation of hourly
564 groundwater levels in a coastal aquifer system of the Venice lagoon. *Engineering Applications of*
565 *Artificial Intelligence*, 25, 1670-1676.
566 Trichakis I.C., Nikolos I.K., Karatzas G.P., 2011. Artificial neural network (ANN) based modeling
567 for karstic groundwater level simulation. *Water Resources Management*, 25, 4, 1143-1152.
568 Valdes D., Dupont J-P, Laignel B., Slimani S., Delbart C., 2014. Infiltration processes in karstic
569 chalk investigated through a spatial analysis of the geochemical properties of the groundwater: The
570 effect of the superficial layer of clay-with-flints, *Journal of Hydrology*, 519, 23-33.
571 Zacharias, S., Bogena, H., Samaniego, L., Mauder, M., Fuß, R., Pütz, T., Frenzel, M., Schwank,
572 M., Baessler, C., Butterbach-Bahl, K., Bens, O., Borg, E., Brauer, A., Dietrich, P., Hajnsek, I., Helle,
573 G., Kiese, R., Kunstmann, H., Klotz, S., Munch, J.C., Papen, H., Priesack, E., Schmid, H.P.,
574 Steinbrecher, R., Rosenbaum, U., Teutsch, G. and Vereecken, H., 2011. A Network of Terrestrial
575 Environmental Observatories in Germany. *Vadose Zone Journal*, 10, 955-973.
576 Zhang, J., Zhu, Y., Zhang, X., Ye, M., Yang, J., 2018. Developing a Long Short-Term Memory
577 (LSTM) based Model for Predicting Water Table Depth in Agricultural Areas. *Journal of Hydrology*,
578 561, 918-929.