MLRA2021 forecasting event
Machine learning prediction event for the international conference in "Machine learning & Risk assessment in geoengineering", Wroclaw, Poland, 25-27 October 2021

MLRA2021 groundwater time-series forecasting via LSTM

Mingliang Zhou[*1], Linhan Ouyang[1], Cong Nie[1], Yu Yu[1], Dongming Zhang[1], and Hongwei Huang[1]

[1] Department of Geotechnical Engineering, Tongji University, Shanghai, China.

[*]Presenting author (email: zhoum@tongji.edu.cn)

1 INTRODUCTION

The machine learning competition invites geotechnical engineers to conduct groundwater time-series forecasting based on actual site project data. The combined railway and road project called FRE16 is in the southeastern part of Norway, which consists of long tunnels, open pits, bridges, and railway-stations. The construction of the 40 km long megaproject is planned to launch in 2022, which can impact the nearby environment.

The construction activities can influence the groundwater level and consequently detrimental effects on nature and buildings in a broad zone along with the project. As a result, monitoring and forecasting the natural groundwater level is vital, un-affected by construction activities. The actual measured values in the construction period can then be compared with the predicted values.

This study utilizes the installed monitoring system, which has 240 sensors along the planned line for the FRE16 project. The sensor measurements in the file train.csv can be used to train a time-series model. The long short-term memory (LSTM) network is adopted to learn the provided time-series data and infer future groundwater measurements. In addition, the geolocations and the physical relations of the sensors corresponding to the six target sensors are considered extra input information to aid the LSTM network training. Hence, a physics-informed LSTM is established in this study to make time-series forecasting of the groundwater measurements.

2 DATASET PREPARATION

Among the 240 sensors, six sensor measurements are obtained, namely temperature, precipitation, air pressure, pore pressure, the water level in the core, and water level in the lake. It is found that only a tiny portion of the sensor data is provided in the Kaggle platform at the interval from 2017-03-08 to 2018-12-01. This can be caused by the malfunction of some sensors or a data-logging system. Therefore, the data at this period is not used for training the LSTM model, and only the data provided for the interval from 2018-12-01 to 2020-06-29 is used.

Nevertheless, the adopted data from 2018-12-01 to 2020-06-29 also has some missing measurements by malfunction sensors and the outlier measurements with a Z-score larger than 3. By adopting the adjacent principle, we fill the missing measurements with the mean of nearby data and fill the outlier measurements with the largest data value nearby.

The sensors' geolocation locations (X, Y, Z coordinates) relative to the six sensors are also essential. A threshold parameter R is selected, and the sensors within the radius of R are chosen as the relevant sensor data for input training. Among the three available feature-data

types, the precipitation has a much stronger relationship with groundwater level than the other two feature-data, namely temperature and air pressure. As a result, the weight of each feature data is assigned separately to achieve the best performance.

It should also be noted that the scale of groundwater fluctuation can vary among different months/seasons of the year. Therefore, the scales should be adjusted best to reflect the scale fluctuation of a specific period.

3 RESULTS & CONCLUSION

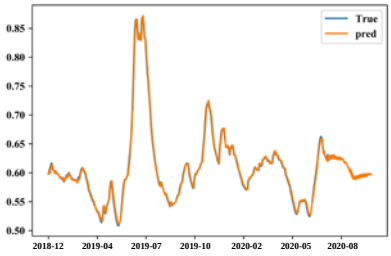
This study adopted the LSTM network to train and test based on the prepared training and testing dataset. The input data is processed from the procedure described in Section 2. The model parameters are tuned to minimize the RMSE value at the training stage to achieve the best forecasting performance. The adopted model hyper-parameters are listed in Table 1, and the corresponding loss values of the six target sensors are also presented.

In addition, the threshold parameter R, the weight of the feature data, and the scales at a different time of the year are also tested with different values to achieve the best loss value. The training and predicting results of the six sensors are shown in Fig 1. The established LSTM model can predict the train stage time series of the ground water and water level very well. The forecasting data suggests a reasonable trend, and the loss value based on the Kaggle evaluation also indicates reliable performance. Due to the small range of the pore pressure measurements, the established model did not perform so well for the pore pressure time-series.

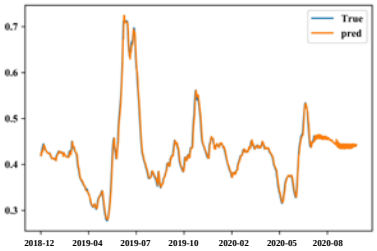
Overall, this physical-informed LSTM model combines sensor geolocations, physical relationships, and machine learning to make the groundwater time-series forecasting results more creditable and explainable. For the coming construction actives in 2022, this model can infer construction influence on groundwater level and inform the site engineers to carry out the correct actions to mitigate the effect.

Table 1 Hyper parameters of the established LSTM model and the loss values

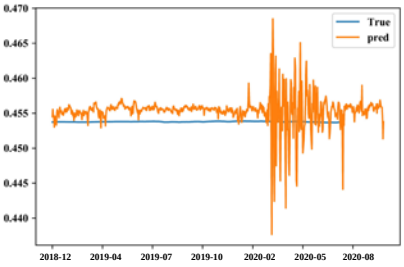
Parameter	**Ground water**		**Pore pressure**		**Water level**	
Sensor label	4C09006	4C11009	5G09018Z_A	K02_A	VS_ABBOR	VS_Kroksund
Epoch	3000					
LSTM layers	6					
Batch size	72					
Moving time window	1 day					
Training Loss	1.36e-05	2.87e-05	5.31e-07	1.89e-07	2.74e-06	1.66e-06
Validation loss	5.7e-06	1.55e-05	7.87e-06	1.88e-06	9.73e-07	1.16e-05



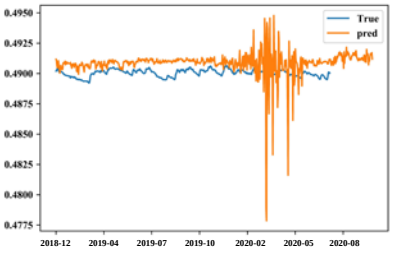
(a) ground water 4C09006



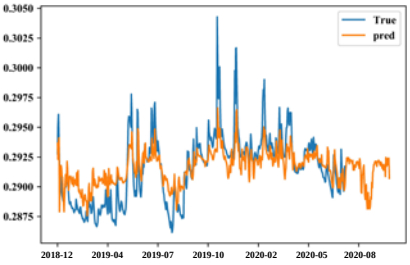
(b) ground water 4C11009



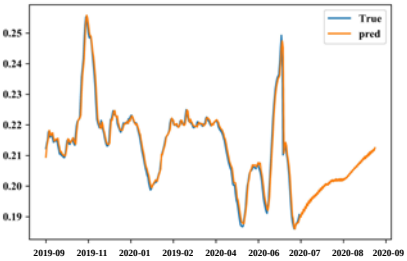
(c) pore pressure 5G09018Z_A



(d) pore pressure K02_A



(e) water level vs_ABBOR



(f) water level vs_Kroksund

Figure 1 Sensor measured ‘true’ data VS the predicted data from the established LSTM model of the 6 sensors