Groundwater time series analysis with clustering and LSTMs for sporadic monitoring datasets

Abstract

Groundwater is an essential source of freshwater in many regions of the world, and it is in constant flux due to evolving anthropogenic influences and changing climate conditions. Careful regional monitoring and modelling of this resource are imperative to ensure its sustainability into the future. However, traditional mathematical models of groundwater systems that represent the physical processes through a series of differential equations require a great deal of information about the hydrogeological system which is often unknown or very difficult to collect. Therefore, scientists interested in groundwater analyses are increasingly turning to machine learning techniques, and deep learning in particular, to efficiently analyse the available data, especially when detailed knowledge of the physical system and its anthropogenic influences is scarce. But groundwater time series often do not provide the large datasets preferred for machine learning and the extremely sporadic measurements can make innovative methods such as the LSTM, a deep learning network specifically for time series analysis, inaccessible for their analysis. However, current research in the field of machine learning is supporting the merging of many simultaneous time series, building models incorporating information from groups of time series to produce predictions on individual time series. A case study from the Namoi River area of Australia (a dry and over-exploited groundwater region characterised by complex hydrogeological processes), is discussed in which multiple time series are grouped together enabling groundwater level predictions with the LSTM in places where the LSTM could not be used on the individual time series. The information gained by combining these time series through the blending of unsupervised and supervised learning techniques is illustrated; no detailed knowledge of the hydrogeological conditions is required. This method facilitates the use of advanced machine learning methods on data that otherwise would not support the use of these models, while also providing the opportunity for a visual analysis of patterns of historic groundwater levels. The method could be used to study groundwater patterns and temporal trends in any region of the world where there is a network of monitoring bores but scant information on the subsurface geology required for traditional models.

Key words: hydrology, LSTM, deep learning, clustering, self-organizing map, environmental monitoring data, missing data

Introduction

Much of the world’s population depends on groundwater as their primary source of freshwater. Yet groundwater resources are constantly fluctuating due to anthropogenic influences and changing climate systems. It is important to monitor and model these changes in planning for a sustainable future, especially in water-stressed regions. Traditional groundwater models involve complex systems of differential equations that mimic the physical processes that contribute to groundwater levels. But the amount of detailed information needed to accurately replicate an area with a physically-based model is extensive, and is usually difficult, time-consuming and expensive to obtain. Many areas do not have the information required to build these mathematical models without making many assumptions about conditions such as hydraulic conductivity, subsurface geological structures, aquifer size and connectivity, and characteristics of the soil and vegetation.

Researchers involved in groundwater modelling and predictions are increasingly turning to data-driven machine learning techniques such as neural networks, and deep learning methods in particular, since these methods are able to extract features and relationships from the available data without requiring expert knowledge of the system (Shen, 2018; Reichstein et al., 2019, Lee et al., 2019). The long short-term memory algorithm, LSTM (Hochreiter & Schmidhuber, 2017), is a deep-learning network specifically for time series analysis that is quickly gaining popularity in recent years in hydrologic modelling due to its ability to make accurate predictions on time series through the learning of temporal patterns and inter-variable relationships. However, groundwater time series often do not provide the large datasets preferred for machine learning and the extremely sporadic measurements can make innovative methods such as the LSTM inaccessible for their analysis.

An area of the Namoi River valley of eastern Australia is considered in this study. This is a very dry area that has been overexploited in terms of groundwater extractions (Prosser et al., 2011; Rassam et al., 2013; Welsh, 2014). The subsurface hydrogeology consists of multiple aquifers, and the response of groundwater levels to external influences is complex. The available time series from hundreds of bores in the area are characterised by sporadic measurements and differing lengths, and so this data does not support the use of the LSTM on the individual bores’ time series.

The machine learning field has recently been producing literature on techniques that merge information from multiple concurrent time series in order to enable the sharing of pattern information between the time series (Bandara et al., 2018; Montero-Manso et al., 2020). A global model is produced based on the combined data set which is able to predict information for each individual time series; rather than assuming each time series comes from a different data-generating process, combining them allows information to be revealed that may be common across the system.

The objectives of this work are to use the time series from multiple locations combined with readily available climate data, and no detailed hydrogeological input or assumptions, to understand historic groundwater conditions in the region and to predict future conditions. A two-stage methodology blends visual analytics with deep learning methods: first, an unsupervised clustering is performed with a self-organising map (SOM, Kohonen, 1990) to determine similar temporal patterns in the time series; and then supervised modelling with the LSTM produces predictions, for each group of similar time series, of water levels responding to certain climate and extraction conditions.

Data and Methods

The time series used in this study include groundwater levels, rainfall, evapotranspiration, river discharge, and annual groundwater extraction volumes from January 1974 to December 2018. The groundwater and surface flow data are from the WaterNSW website [w1], a sample of which are shown in **Figure 1**, and the rainfall and evapotranspiration data are from SILO [w2].

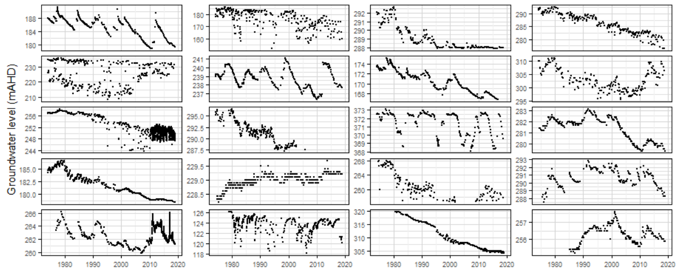

Figure 1: Sample water level time series for a selection of groundwater bores (1974-2018). It can be seen that the temporal patterns are similar between some bores and differ greatly between others.

The self-organizing map, an unsupervised learning algorithm from the family of neural networks, is used for pattern extraction, clustering and data visualisation. Prevalent patterns in the data set are identified and each data item is allocated to its best match amongst these patterns. The SOM is resilient to high levels of missing data, and is therefore a popular choice for analysing environmental data. Because the algorithm is unsupervised, only the bore water level data are used at this point without having to include climate and extraction data. Due to their proximity, the bores must be experiencing relatively similar ambient conditions at a given time. The resulting clusters therefore represent bores that respond similarly to external conditions, even though these have not been specified at this stage.

The LSTM is an updated form of the recurrent neural network that allows for much longer time series to be analysed. The LSTM 'looks back' over the data for a specified amount of time, cycling sequentially through the observations and retaining important information over time. External forcing data can be incorporated while determining multi-scale temporal dependencies. In training an LSTM for each group of bores, monthly averages of rainfall and evaporation from 6 climate stations, surface flows from 5 rivers and 1 dam, and annual extraction data at 60 pumping locations are used as predictors. The representative time series generated by the SOM, for each group of bores with similar temporal patterns, are used as the response variables.

Results

The SOM output map is shown on the left of **Figure 2**. Sixteen grey patterns of monthly water levels can be seen, which form the set of 'representative' time series, one for each of the identified clusters. These represent the sixteen most prevalent temporal patterns determined to exist in the data. Coloured smoothers have been added to indicate the overall temporal pattern of each cluster. These patterns are arranged on the grid so that patterns near each other are more similar than those further apart. The purple and dark blue patterns in the upper left show water levels decreasing over time, the green patterns in the middle descend but then start to even out or slightly increase, the yellowish-green in the lower right is relatively constant, and the yellow pattern in the lower left indicates water levels increasing steadily over time.

Each individual bore time series is then matched to the pattern that it fits best. On the right of **Figure 2**, we can see the list of bores and their recorded data for a selection of SOM clusters. At node 1, there is a group of 9 bores with water levels that are decreasing steadily over time. At node 12, water levels are generally decreasing, with annual drawdowns increasing greatly around the middle of the study period for this group of 11 bores. At node 13, we see 10 bores that have generally increasing water levels with similar multi-year fluctuations.

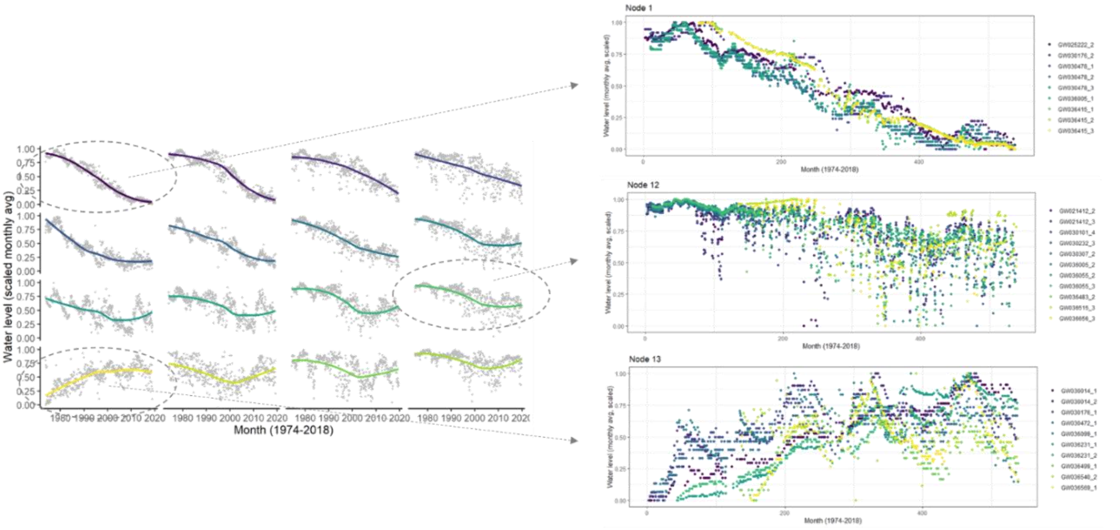

Figure 2: SOM output showing 16 prevalent time series patterns in the data set (grey dots) with smoothers (coloured lines) (left); and the individual time series that are matched to each pattern (right).

Now that it is known which cluster each bore time series is assigned to, they can be visualised in geographic space including this information as colour. **Figure 3** shows water level time series as columns, for bores in the deep aquifer of the Upper Namoi region, coloured corresponding to the patterns in **Figure 2**. The column widths correspond to water level, and change as water levels increase and deplete over time, with the base of the columns at January 1974 and the tops at December 2018. Narrower points in the columns indicating months when water levels are low.

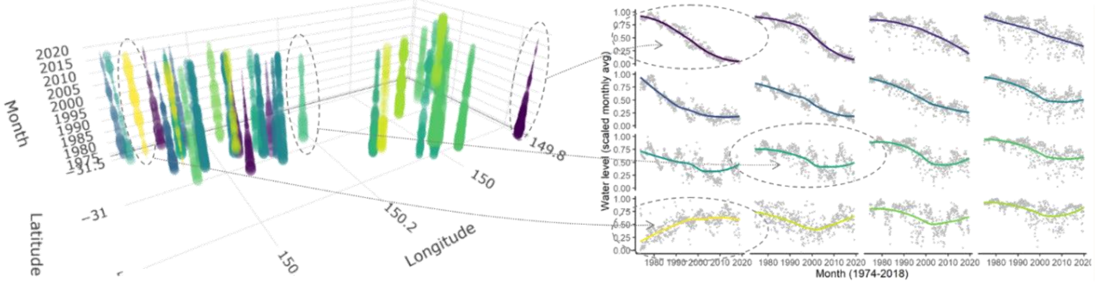

Figure 3: Monthly water level time series for Upper Namoi deep bores plotted in latitude/longitude with diameter of column indicating amount of water present at each month; colours correspond to SOM clusters.

Figure 4 shows the location and cluster membership of the deep aquifer bores on a map of the region; the legend includes a reminder of the general temporal water level pattern for each cluster.

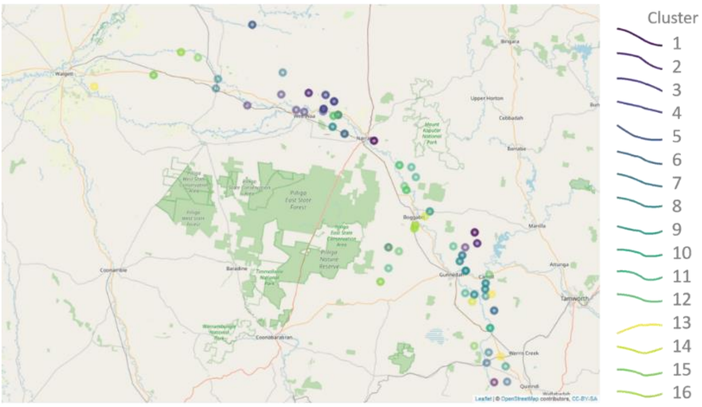

Figure 4: Upper and Lower Namoi deep bores, coloured by cluster membership. The historical water level trends can be seen in the context of the geography of the region.

As well as a visual insight into the historic patterns of groundwater levels in the region, the clusters of bores now provide additional information that can be used in an LSTM model for prediction. A model created with the representative time series from each cluster can be used to model the processes for each group of bores. Predictors of rainfall, evaporation, surface flow and extractions are now added. The representative time series of each cluster becomes the response variable to this set of predictors. A separate LSTM is created for each cluster, but the set of predictors is the same for all. As an example, the left panel of **Figure 5** shows the representative time series for cluster 15 in grey, with the measurement data from the member individual bores overlaid in colour. The LSTM prediction is shown in red on the right panel, again with the cluster representative time series in grey.

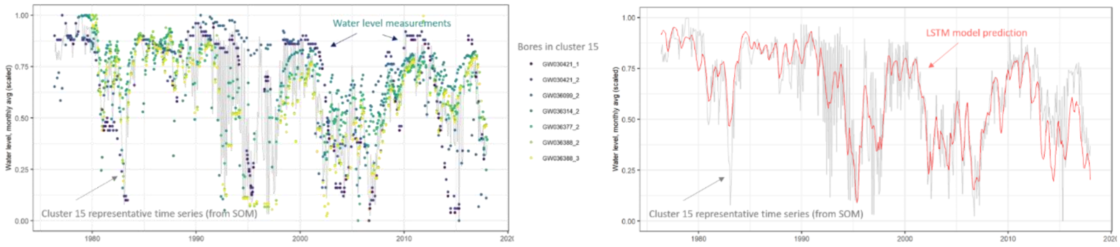

Figure 5: Cluster 15 representative time series (grey) and a) observed data at each associated bore, b) LSTM prediction

After training the models on the representative time series of each cluster, the predictions can be compared to the measurements from individual bores within the clusters. In **Figure 6**, the cluster representative time series line is removed and measurement points for two individual bores from within this cluster are overlayed. In doing so, it can be seen that the predictions track the infrequent measurements at these bores, for which we were unable to create individual LSTMs due to uneven measurement intervals. Though one bore is in the Upper Namoi region and the other in the Lower Namoi region, the LSTM cluster prediction generally follows the data points from both measurement time series, even though the water levels measurements come from different aquifers.

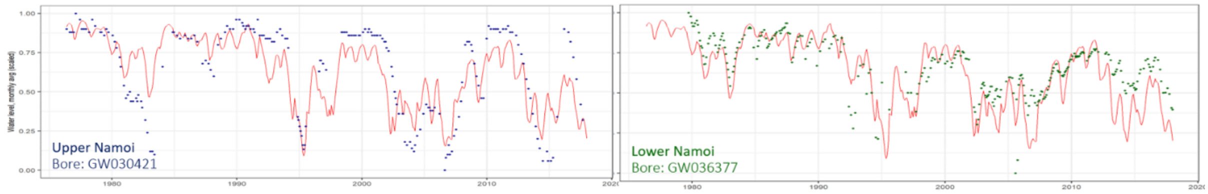

Figure 6: LSTM prediction (red) for Cluster 15 is shown with measurement data (blue and green dots) from individual bores in two different regions

Conclusion

A data-driven temporal analysis has been performed of groundwater levels in a dry and over-exploited region of Australia which is characterised by complex, and not comprehensively defined, hydrogeological processes. Information from multiple time series was merged, allowing prevalent patterns in the overall data set to be used to make predictions of groundwater levels at individual locations, with no expert knowledge of subsurface conditions required. A set of general models has been created that can represent bores from different areas and different aquifers, as long as the groundwater levels in those areas respond similarly to the existing climate and pumping conditions of the regions. Visual analytics is an important tool in aiding water practitioners in the decision making process, and this method also provides the opportunity for visual exploration of historical groundwater patterns. Combining sporadic time series from multiple bores to facilitate the use of state-of-the-art data-driven time series modelling, this method may be useful in regions of the world where groundwater management is needed but detailed hydrogeologic knowledge is sparse.

References

[w1] https://realtimedata.waternsw.com.au/

[w2] https://www.longpaddock.qld.gov.au/silo/

Bandara, K., et al. (2020). "Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach." Expert systems with applications 140: 112896.

Hochreiter, S. and J. Schmidhuber (1997). "LSTM can solve hard long time lag problems." Advances in neural information processing systems: 473-479.

Kohonen, T. (1990). "The self-organizing map." Proceedings of the IEEE 78(9): 1464-1480.

Lee, S., et al. (2019). "Using artificial neural network models for groundwater level forecasting and assessment of the relative impacts of influencing factors." Hydrogeology Journal 27(2): 567-579.

Montero-Manso, P. and R. J. Hyndman (2020). "Principles and algorithms for forecasting groups of time series: Locality and globality." arXiv preprint arXiv: 2008.00444.

Prosser, I. P. (2011). Water: science and solutions for Australia, CSIRO.

Rassam, D. W., et al. (2013). "Accounting for surface–groundwater interactions and their uncertainty in river and groundwater models: a case study in the Namoi River, Australia." Environmental modelling & software 50: 108-119.

Reichstein, M., et al. (2019). "Deep learning and process understanding for data-driven Earth system science." Nature 566(7743): 195-204.

Shen, C. (2018). "A transdisciplinary review of deep learning research and its relevance for water resources scientists." Water Resources Research 54(11): 8558-8593.

Welsh, W., et al. (2014). Context statement for the Namoi subregion. Product 1.1 from the Northern Inland Catchments Bioregional Assessment. Department of the Environment, Bureau of Meteorology, CSIRO and Geoscience Australia.