Workshop on Visualisation in Environmental Sciences (EnvirVis) (2018)
R. Bujack and K. Rink and D. Zeckzer and S. Jänicke (Editors)

Predict Saturated Thickness using TensorBoard Visualization

Vinh The Nguyen, Tommy Dang, and Fang Jin

Computer Science Department
Texas Tech University, Lubbock, USA

Abstract

Water plays a critical role in our living and manufacturing activities. The continuously growing exploitation of water over the aquifer poses a risk for over-extraction and pollution, leading to many negative effects on land irrigation. Therefore, predicting aquifer water level accurately is urgently important, which can help us prepare water demands ahead of time. In this study, we employ the Long-Short Term Memory (LSTM) model to predict the saturated thickness of an aquifer in the Southern High Plains Aquifer System in Texas, and exploit TensorBoard as a guide for model configurations. The Root Mean Squared Error of this study shows that the LSTM model can provide a good prediction capability using multiple data sources, and provides a good visualization tool to help us understand and evaluate the model configuration.

CCS Concepts

• ***Information systems*** → *Information systems applications;* •***Information systems*** → *Information retrieval;* •***Information systems applications*** → *Data mining;*

1. Introduction

Water is the basic element that human relies on for all living and manufacturing activities. As rainfalls are not equally distributed in the world, surface water does not meet the demand to sustain many areas. In this case, it is necessary to turn to groundwater found in aquifers to support daily activities. The Ogallala Aquifer [GH70] is underground water surrounded by sand, silt, clay and gravel, underlying approximately 175,000 square miles from South Dakotas to the Texas in United States. The Ogallala Aquifer of Texas provides 96 percent of underground water for irrigation and 36 percent for municipal demands and thus plays a critical role to the economical development of this region. Wells drilled in these areas must reach to a certain point to pump up water and this point must be in the saturated thickness, which is the distance between the water table and base of the aquifer. Due to the excessive use of water, the saturated thickness has declined consistently through time [GMP11] although there are numerous recommendations of using the Ogallala Aquifer, including drilling new wells, over-drafting, reallocating supplies or developing well fields [GMP11]. From this point of view, there is a need to have good water management strategies for proper water management strategies for a sustainable use of the aquifer system. One necessary requirement is to continuously monitor groundwater levels [DNKU17].

There are number of studies conducted to address the challenges of the Ogallala Aquifer Systems and provide some projections [BM71, BM79, DRM01, McA84, SBY*13]. However, most of these studies focus on economy impacts, specific to other regions or indicators that are not relevant to current time due to the changing of economical growth or expansion of non-agriculture areas. As the saturated thickness being depleted, the purpose of our study is to predict the saturated thickness as the first step for water management. In addition, this study also provides an indicator for non-experts in machine learning to select suitable configurations for the LSTM model.

The key contributions of this paper thus are:

- It employs the Long-Short Term Memory (LSTM) model to predict the saturated thickness of eight counties in Texas. This model can be extended to any other counties given sufficient data.
- It integrates the TensorBoard visualization which enables users to analyze and optimize model configurations
- It reports the performance of the trained model on eight data-sets.

The rest of this paper is organized as follows: In Section 2, a summary of existing work is presented. Section 3 describes the data-sets along with model development based on the TensorFlow framework. Performance and result are discussed in Section 4. Finally, the conclusion and future work are represented in Section 5.

2. Related Work

There are plenty of water forecasting studied and reported in the literature. In this section, we have no intention to exhaustively review all of them. Instead, we discuss the most relevant work to our study.

Currently, there are only a few studies conducted to predict the

DOI: 10.2312/envirvis.20181135

saturated thickness of the Ogallala Systems in Texas. A complete study made by Dutton et al. [DRM01] in 2001, they created a conceptual model that was capable of predicting underground water levels of 18 counties in Texas by 2050. However, this study was based on the assumptions that pumping remains constant until a well is depleted and projected pumping rates were based on the continuation of agriculture and economic development. These indicators are unstable since a reported by RN Wilkins et al. [WSF*09] in 2009 indicated that approximately 100,000 acres of Texas working lands were converted to non-agricultural uses from 2007 to 2012.

Steward et al. [SBY*13] proposed an integrated method for forecasting groundwater depletion by 2110. The study found that nearly 70 percent of underwater will be depleted in the next 50 years given the current trends in Kansas. A logistic function was developed based on the dimensionless saturated thickness to approximate groundwater level over time.

This paper tries to solve the same problem but with a different approach, it provides a guideline for hydrologists to look into the *blackbox* of model training and choose choose an optimal configuration.

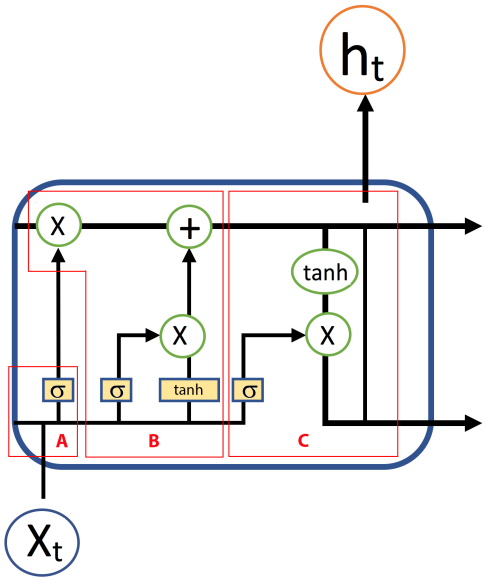

Figure 1: *Long Short-Term Neural Network architecture*

3. Methods

The data set in this study is a time series data, that is, the saturated thickness is observed, recorded and indexed in time order. Time series prediction is to have a model to predict the future values based on the previously observed values. The most popular model for time series prediction is the Autoregressive Integrated Moving Average (ARIMA) model. However, this model has two main drawbacks because of its assumptions, that is, there is a linear relationships between independent and dependent variables and a constant standard deviation in errors in the model over time [KPSR14]. Real world data often does not often satisfy these assumptions as shown in our study data set in Figure 3. The Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) model [Bol86] can be employed to elaborated these assumptions, however optimizing the GARCH model parameters is a challenging task. In recent years, deep learning has gained its popularity to address the existing challenges of time series prediction, particularly the Long Short Term Memory (LSTM) model proposed by Sepp Hochreiter and Jurgen Schmidhuber [HS97]. The LSTM model is basically a Recurrent Neural Network, it is capable of predicting future values based on not only previous values but also long past values in sequence. Cell state is the key to LSTM which is the horizontal line running through the top of the diagram as depicted in Figure 1. Information in the cell state can be added or removed by three operational gates.

- **Forget gate (Figure 1–A):** This sigmoid layer decides what information will be thrown away through a function: $f_t = \sigma\left(W_f\left[h_{t-1},x_t\right]+b_f\right)$. This function gives output between 0 and 1. Value of 0 means *completely get rid of this* while value of 1 indicates *completely keep this information*
- **Input gate (Figure 1–B):** The previous output and the new input are taken by this gate and passed through another sigmoid layer. This gate also returns an output between 0 and 1 by a function $i_t = \sigma\left(W_i\left[h_{t-1},x_t\right]+b_f\right)$. A vector of new candidate values is created $\tilde{C} = tanh\left(W_c\left[h_{t-1},x_t\right]+b_c\right)$, then combined with the value of the input gate $(i_t * \tilde{C})$ and old state $(C_{t-1} * f_t)$ to decide how much to update each state value. $C_t = C_{t-1} * f_t + i_t * \tilde{C}$
- **Output gate (Figure 1–C):** This gate decides how much of the internal state will be passed to the output. First, for deciding what parts of the cell state will be output we run a sigmoid layer: $o_t = \sigma\left(W_o\left[h_{t-1},x_t\right]+b_o\right)$. Then, the cell state is put through **tanh** and multiplied with the output of the sigmoid gate $h_t = o_t * tanh(C_t)$

3.1. TensorFlow Architecture

In 2011, The Google Brain [Bra11] project was started to explore the use of very large scale deep neural networks. As a result of this project, TensorFlow [AAB*16] is the second–generation machine learning system, which uses data flow graphs (Figure 2) to build models. Compared to its predecessor DistBelief [DCM*12], this system is more flexible, scalable, and better performed, especially it supports a wide range of models for training on a variety of heterogeneous hardware platforms. In Figure 2, the instantiation of an operation is represented by a node which has zero or more inputs/output. The edges (or paths) of the graph allow the data to flow from node to node. Value that flows along the edges is called *tensor*, which is a multidimensional array. Because of the dynamically sized data arrays, it is possible to create almost any type of data flow graph.TensorBoard has special features to view the machine learning model and its ability to evaluate the performance of the models with desired metrics. This paper employs the TensorBoard framework to analyze the neural network model.

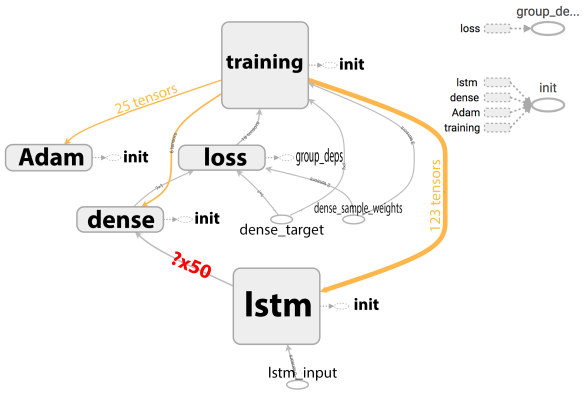

Figure 2: *Example of TensorFlow graph, some parts of the graph are enlarged for better visibility*

3.2. Data-set and data pre-processing

Data Description: The data for this study was retrieved from two different sources (Water Resources Center [Uni17] and Water Data for Texas [DWM17]), which contains 9 features (ID of well, longitude of well, latitude of well, county, month of measuring the saturated thickness, day of measuring the saturated thickness, year of measuring the saturated thickness, water elevation, and the saturated thickness). Figure 3 provides a brief trend of wells data over eight counties. Observations were mostly collected from 1/2002 to 9/2016. The two last county data (Swisher and Deaf Smith) were collected from 12/2002 to 9/2016.

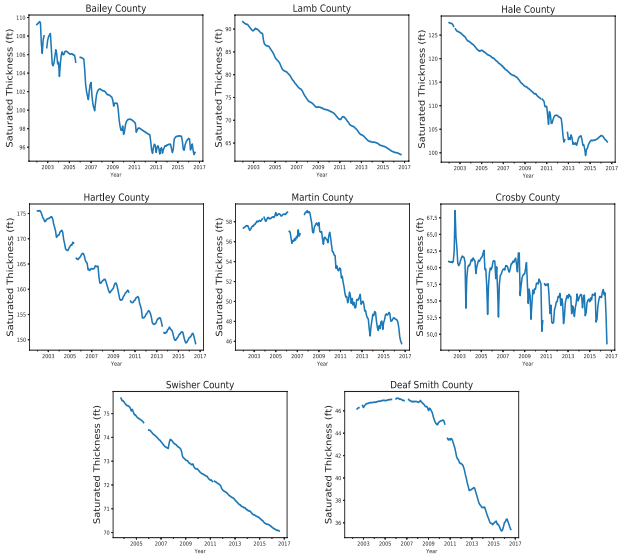

Figure 3: *Data trends of 8 counties from 2002 to 2016*

Handling missing data: Some data are missed in some months, as depicted by a discontinued line in Figure 3. Linear regression method [CCWA13] is applied to impute the missing values.

As shown in Figure 3, there is trending down in the time series data, so normalization is applied for the entire dataset before training. Supervised learning technique is used for predicting the saturated thickness at the current month (t) given the water level, water elevation and saturated thickness measurement at the previous month (t-1).

3.3. Model evaluation

Let y_i denotes the i^{th} observation value, $\hat{y}_i$ is the corresponding predicted value and n is the number of observations. The predicted error is measured by $e_i = y_i - \hat{y}_i$. We use the two most commonly measures for this scale-dependent: Mean Absolute Error (MAE) for training and Root Mean Squared Error (RMSE) to evaluate the performance of the trained model. Mathematical expression of the two measures are given as below respectively.

$$MAE(y,\hat{y}) = \frac{1}{n}\sum_{n=1}^{n} |y_i - \hat{y}_i| \tag{1}$$

$$RMSE(y,\hat{y}) = \sqrt{\frac{1}{n}\sum_{n=1}^{n} (y_i - \hat{y}_i)^2} \tag{2}$$

4. Experimental Results

4.1. Experiment Setup

For training and testing data. The first 85 % observations are used for training and the remaining 15 % observation are used for testing. We setup the LSTM model with 50 neurons in the first hidden layer, one neuron in the output layer for predicting the saturated thickness. Mean Absolute Error is used as a loss function. For optimization we use the Adam version of the stochastic gradient descent [KB14]. For training the LSTM model, the number of samples per gradient update (batch size) is 72, the number of times iterating over the training data is 100. Additional detail and full implementation of TensorBoard can be found on Github repository at: *https://github.com/Alex-Nguyen/SaturatedThicknessTensorBoard*

4.2. Prediction Performance

It is an arduous task in neural networks to pick up efficient and tight parameters to feed, because neural nets have been considered as a black box [BCR97] in which there is no clear explanation of its behavior. It is essential for scientists to inspect and understand their networks, while TensorBoard is a great web application to support these tasks. Figure 4 provides a brief information of the model performance for each county (each line) based on testing and training data. Predictions and test data are combined and inverted into their original values to calculate the Root Mean Squared Error (RMSE) for the model; it gives an error in the same units as the variable itself. Figure 4 (right) column depicts the error scores of the model for eight counties. Overall, the proposed model performance is good when applying for different datasets with RMSE less than 0.5, the only one exception is with Crosby county (second row) when there is a sudden decrease in the saturated thickness in the last month, but this error score is acceptable compared to the saturated thickness unit.

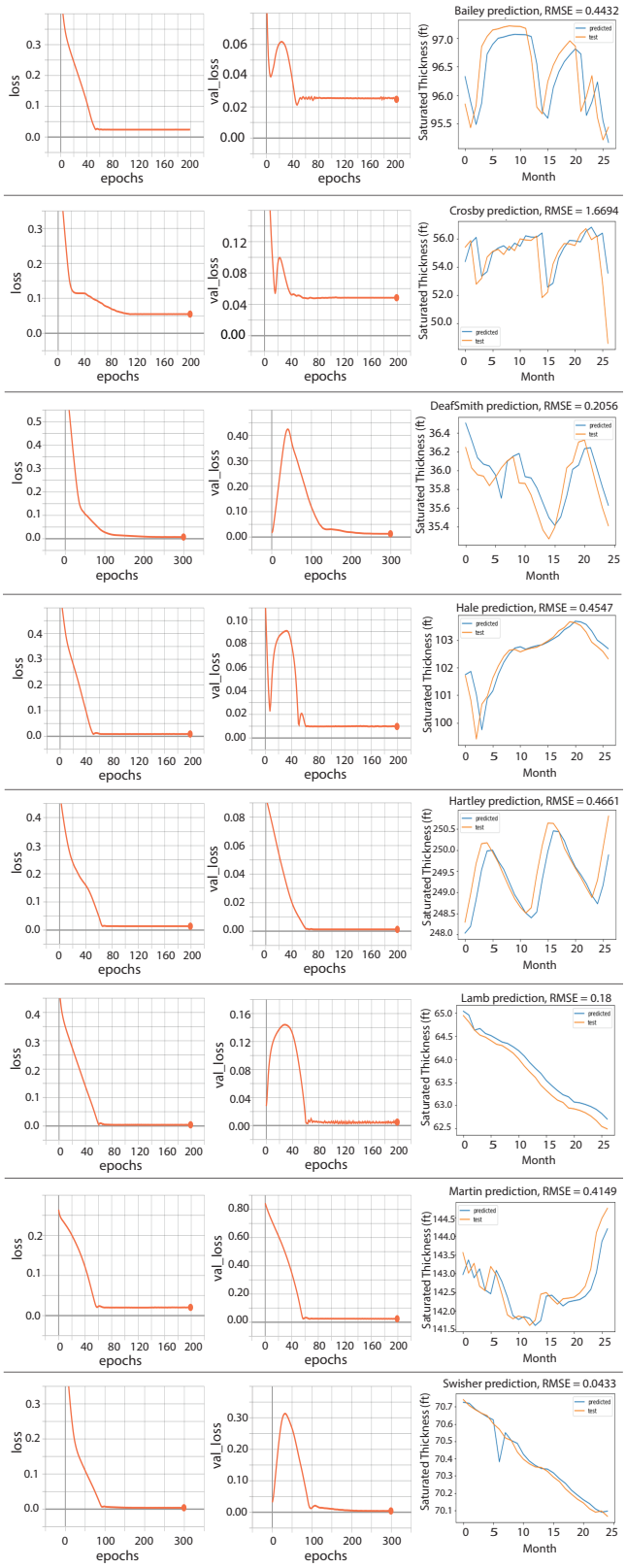

Figure 4: *Prediction performance in terms of RMSE for eight counties. The most left column shows the loss function in the training, the middle column shows the loss function in the test, and the right column shows RMSE value, where the blue curve represents the predicted value, and the orange curve represents the observed value.*

Figure 4 left and middle columns are found to be useful for the configuration of parameters in neural networks. Figure 4 left column shows MAE loss function values during each iteration (or epoch) in the training data while the middle column shows these values in the test data. These two columns provide an indicator whether we should stop training visually. The learning process should stop when the loss begins to stable even with the increase of the number of epochs because the model will not learn any more useful knowledge. In Figure 4, most models start to converge around 100 iterations, except models for Deaf Smith and Swisher counties which need more than 200 epochs to converge. Knowing this information is important because it allows stopping training the model at the early stage to reduce training time or avoid overfitting/underfitting.

4.3. Discussion

While building, constructing and training deep neural network is complex and sometimes confusing to non-experts, TensorBoard serves as a great visualization tool that helps analysts to better understand, debug, and optimize TensorFlow programs by reading from the checkpoint files when TensorFlow variables are saved. When looking at the TensorGraph, as depicted in Figure 2, analysts can check if the proposed neural net is configured correctly (all components should be connected) or debug the tensors flowing along the graph. It can be seen from the Figure 2 that, from *lstm* to *dense*, there is a question mark (?x50) in the TensorGraph, meaning that the dimension is unknown.

Predicting water level is not a new area as there are many researches working on this problems. However, the performance of a studied model mostly depends on selecting parameters to feed into the models or the characteristics of the data. Understanding the black box inside neural networks is still difficult for many researchers, and TensorFlow is one of the suitable tools to unveil this flow of operations. One major problem that we face in this research is the ability to fully understand the whole architecture of TensorFlow since it is still developing and many features are undocumented. Another problem is inadequate data with other wells, one possible approach to overcome this problem is to combine with other resources, such as weather data, water consumption, or social network data, for prediction.

5. Conclusion

In this paper, we applied the LSTM model to predict the saturated thickness of eight counties in Southern High Plains Aquifer System in Texas. Our experiment runs on TensorFlow architecture, and TensorBoard is used to analyze the hidden network layers as well as a guideline for model configurations. The contributions of this paper are two folds: based on the proposed model, it helps scientists, water managers to forecast the saturated thickness in a given area as the first step for water management. Second, it provides a direction for the non-expert in machine learning on how to select appropriate configurations based on model graph visualization. In future work, we are going to focus on real-time prediction when incorporating other information such as weather data and social media.

Acknowledgement

This work was partially supported by the U.S. National Science Foundation under Grant CNS-1737634

References

[AAB*16] ABADI M., AGARWAL A., BARHAM P., BREVDO E., CHEN Z., CITRO C., CORRADO G. S., DAVIS A., DEAN J., DEVIN M., ET AL.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016). 2

[BCR97] BENÍTEZ J. M., CASTRO J. L., REQUENA I.: Are artificial neural networks black boxes? *IEEE Transactions on neural networks 8*, 5 (1997), 1156–1164. 3

[BM71] BELL A. E., MORRISON S.: *Analytical study of the ogallala aquifer in Lubbock County, Texas, projections of saturated thickness, volume of water in storage, pumpage rates, pumping lifts, and well yields*. Tech. rep., Austin, Texas: Texas Department of Water Resources, 1971. 1

[BM79] BELL A. E., MORRISON S.: Analytical study of the ogallala aquifer in carson county, texas, projections of saturated thickness, volume of water in storage, pumpage rates, pumping lifts, and well yields. *Report 242 November 1979. 69 p, 18 Tab, 75 Ref, 24 Map.* (1979). 1

[Bol86] BOLLERSLEV T.: Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics 31*, 3 (1986), 307–327. 2

[Bra11] BRAIN G.: Google brain team's mission, 2011. https://research.google.com/teams/brain/about.html. 2

[CCWA13] COHEN J., COHEN P., WEST S. G., AIKEN L. S.: *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge, 2013. 3

[DCM*12] DEAN J., CORRADO G., MONGA R., CHEN K., DEVIN M., MAO M., SENIOR A., TUCKER P., YANG K., LE Q. V., ET AL.: Large scale distributed deep networks. In *Advances in neural information processing systems* (2012), pp. 1223–1231. 2

[DNKU17] DANG T., NGUYEN L. H., KARIM A., UDDAMERI V.: STOAViz: Visualizing Saturated Thickness of Ogallala Aquifer. In *Workshop on Visualisation in Environmental Sciences (EnvirVis)* (2017), Rink K., Middel A., Zeckzer D., Bujack R., (Eds.), The Eurographics Association. doi:10.2312/envirvis.20171102. 1

[DRM01] DUTTON A. R., REEDY R. C., MACE R. E.: *Saturated thickness in the Ogallala aquifer in the Panhandle Water Planning Area: simulation of 2000 through 2050 withdrawal projections*. Bureau of Economic Geology, University of Texas at Austin, 2001. 1, 2

[DWM17] DWMB: Water data for texas, 2017. https://waterdatafortexas.org/reservoirs/statewide. 2

[GH70] GURU M., HORNE J.: *The Ogallala Aquifer*, vol. 48. WIT Press, 1970. 1

[GMP11] GEORGE P. G., MACE R. E., PETROSSIAN R.: *Aquifers of Texas*. Texas Water Development Board, 2011. 1

[HS97] HOCHREITER S., SCHMIDHUBER J.: Long short-term memory. *Neural computation 9*, 8 (1997), 1735–1780. 2

[KB14] KINGMA D., BA J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 3

[KPSR14] KANE M. J., PRICE N., SCOTCH M., RABINOWITZ P.: Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks. *BMC bioinformatics 15*, 1 (2014), 276. 2

[McA84] MCADA D. P.: Projected water-level declines in the ogallala aquifer in lea county, new mexico, 1984. 1

[SBY*13] STEWARD D. R., BRUSS P. J., YANG X., STAGGENBORG S. A., WELCH S. M., APLEY M. D.: Tapping unsustainable groundwater stores for agricultural production in the high plains aquifer of kansas, projections to 2110. *Proceedings of the National Academy of Sciences 110*, 37 (2013), E3477–E3486. 1, 2

[Uni17] UNIVERSITY T. T.: Water resources center, 2017. https://www.depts.ttu.edu/waterresources/. 2

[WSF*09] WILKINS R. N., SNELGROVE A. G., FITZSIMONS B. C., STEVENER B. M., SKOW K. L., ANDERSON R. E., DUBE A. M.: Texas land trends. *Texas A&M Institute of Renewable Natural Resources, College Station, TX* (2009). 2