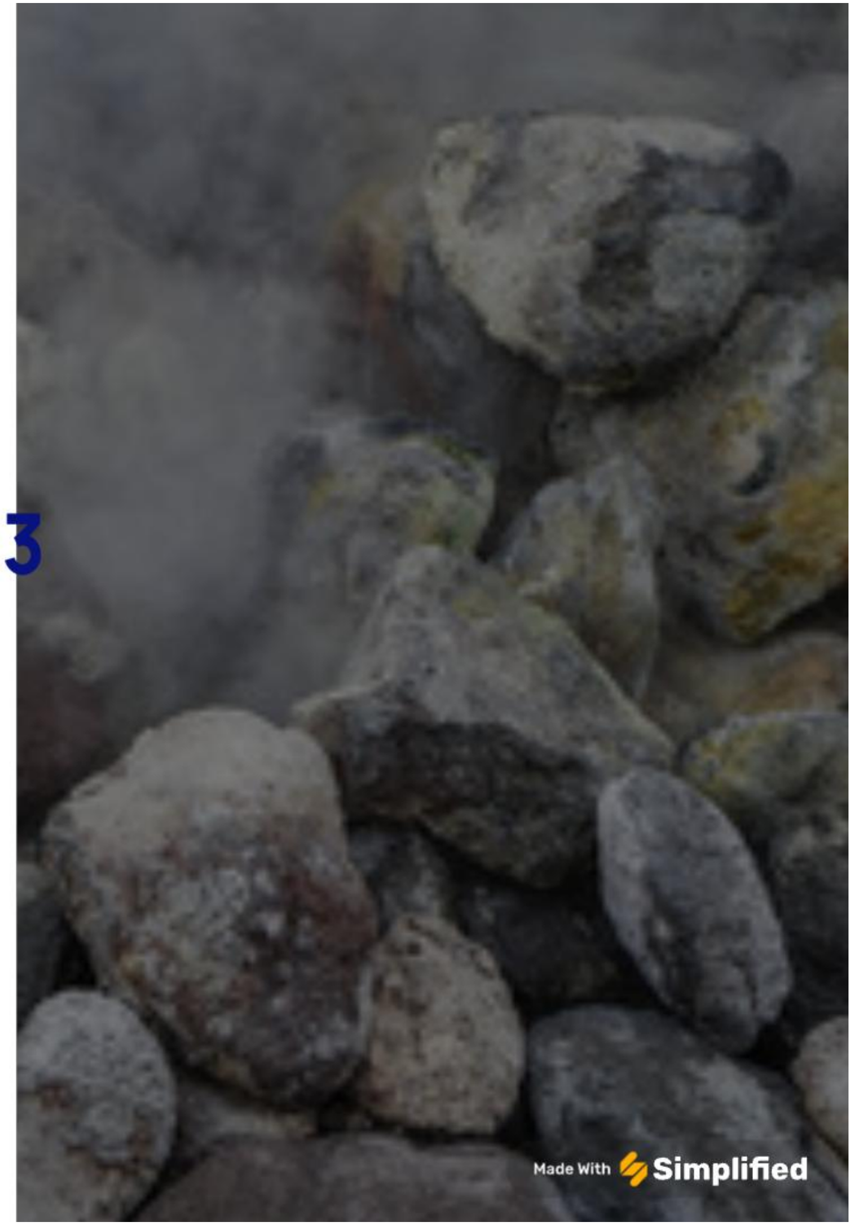
Groundwater Prediction Using Long Short-Term Memory (LSTM)

Name : ESWAR S
Reg.No. : 211521104301
Naan mudhalvan Id: aut2021pitcs293
College Name : Panimalar Institute Of Technology

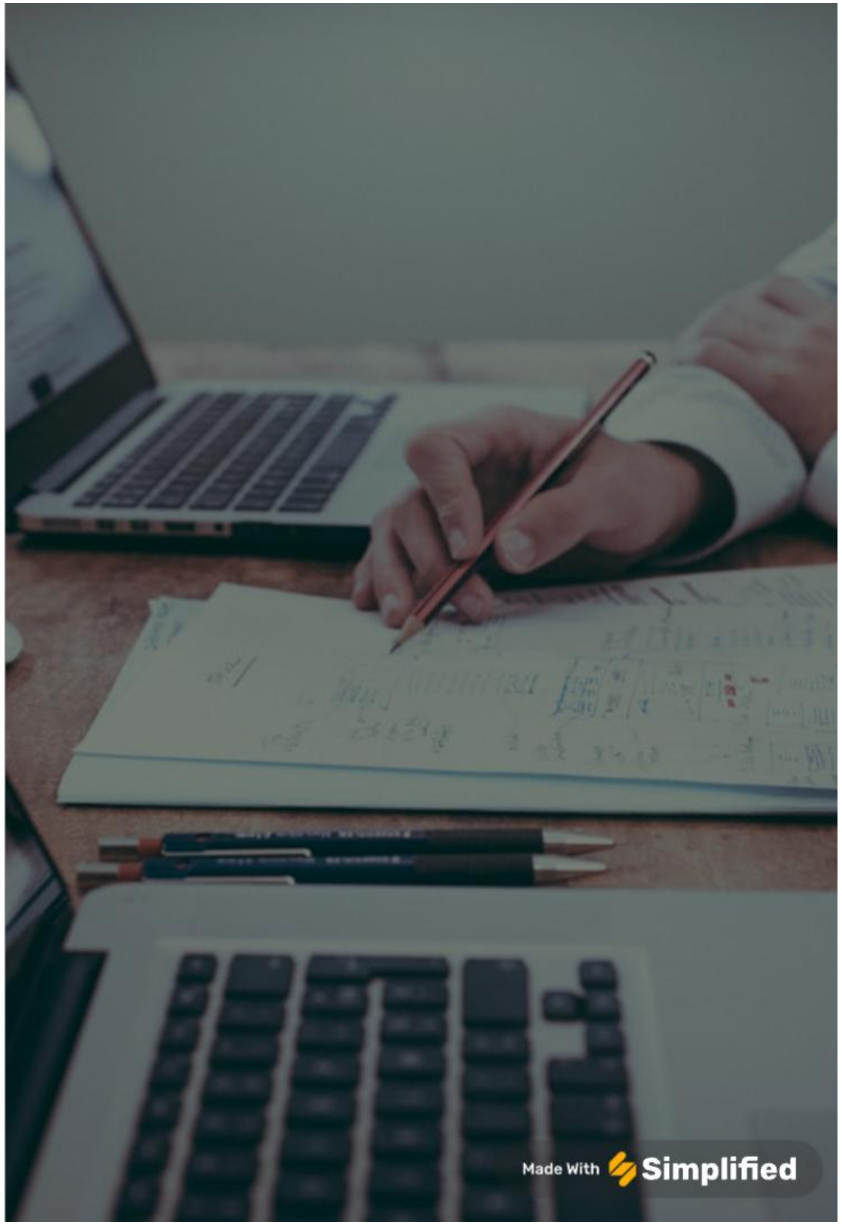


1.Introduction

Water is a general requirement to plant and living organisms on the earth's surface. It is essential for maintainingthe balance of ecology, atmosphere, and natural resources.While natural resources are most important for whole natural life systems on the ground, harmless water must not encompass any harmful chemical materials or living bacteria in concentrations that afect impairment (WHO 2017).Growth and development in the world have led to extensive pollution from rainwater outlets like rivers (UNEP2016). Many factors can afect the chemical, physical, and biotic substances of surface water, for example natural(i.e. rainfall, watershed geography, weather, geology) and anthropogenic activities (i.e. industrial activities, domestic, agricultural run-of) (Mishra et al. 2017; Ewaid et al. 2018; Su et al. 2018).

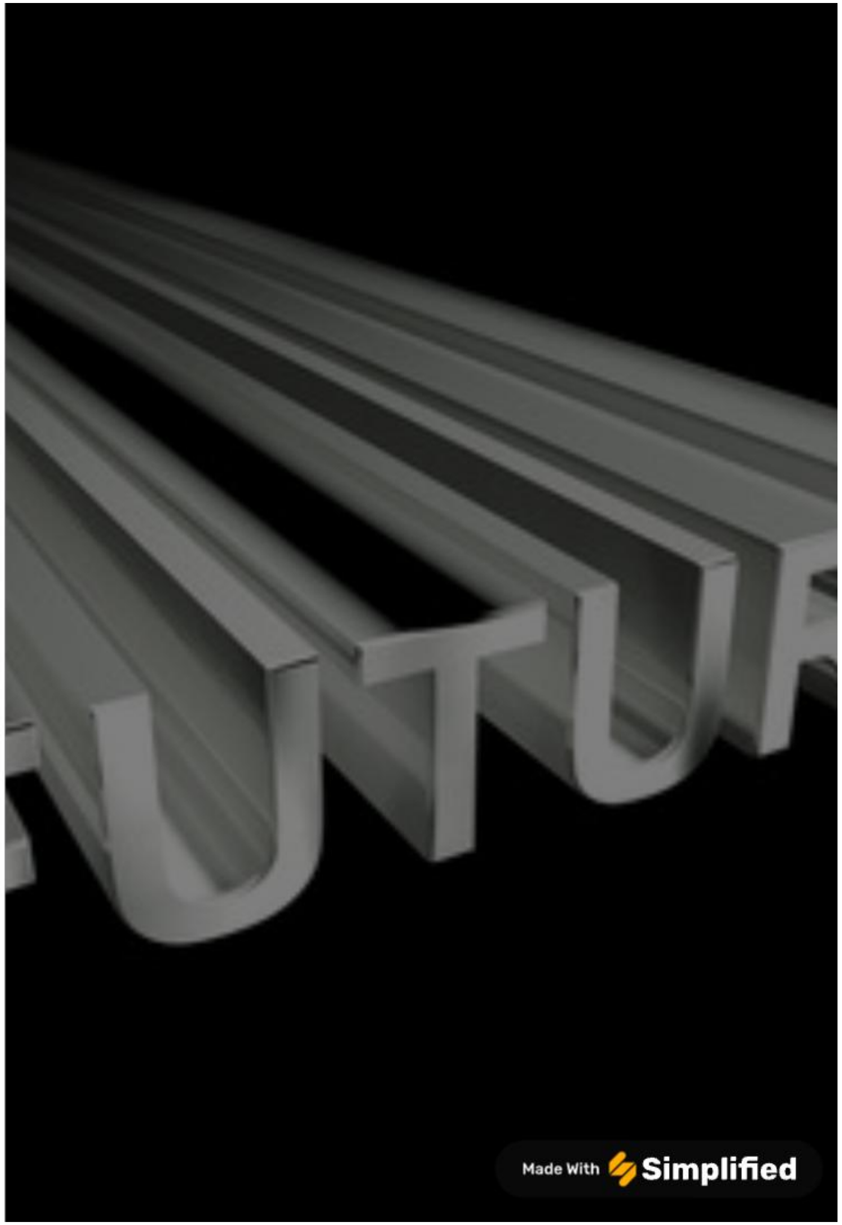
2.Study area

The Akot basin area is situated in the Akot Taluka of Akola district of Maharashtra, between 20°54′30″ and 21°14′35″ N latitudes and between 76°48′ and 77003′E longitudes with 450 sq. km. This study area's minimum and maximum temperatures are 12.6°C and 42.4°C (Fig. 1). The observed annual rainfall is 740 to 860 mm. The deep black is soil found in the southern part of the Akot basin. This soil has a deep, heavy colour with an angular block structure in the sub-surface horizon, medium drained and low to moderate water support. This basin is under the saline water zone because most the groundwater has very highly salted water found within the basin area. In this view, most of the farmers are sufering from so many groundwater quality issues.



3.Methodology

Machine learning models were currently used to estimate most groundwater quality variables precisely and show their efectiveness (Rahgoshay et al. 2018; Ho et al. 2019). One hundred groundwater samples are obtained from this prediction model. The dataset was collected from observation wells within the basin area. We have used 140 water samples for this model. We used 70% of data in training (98 samples) and 15 % of data used in ANN model validation.This research uses 15 % and 30 % data for ANN, MLR, and LSTM models' prediction purposes. This study has developed three machine learning models to predict irrigation water quality parameters, specifcally ANN, LSTM, and MLR models. Therefore, three machine learning models, LSTM, MLR, and ANN, were selected for both prediction scenarios. The frst and second scenarios of MLR, LSTM, and ANN models have been based on all input and reduction variables, respectively

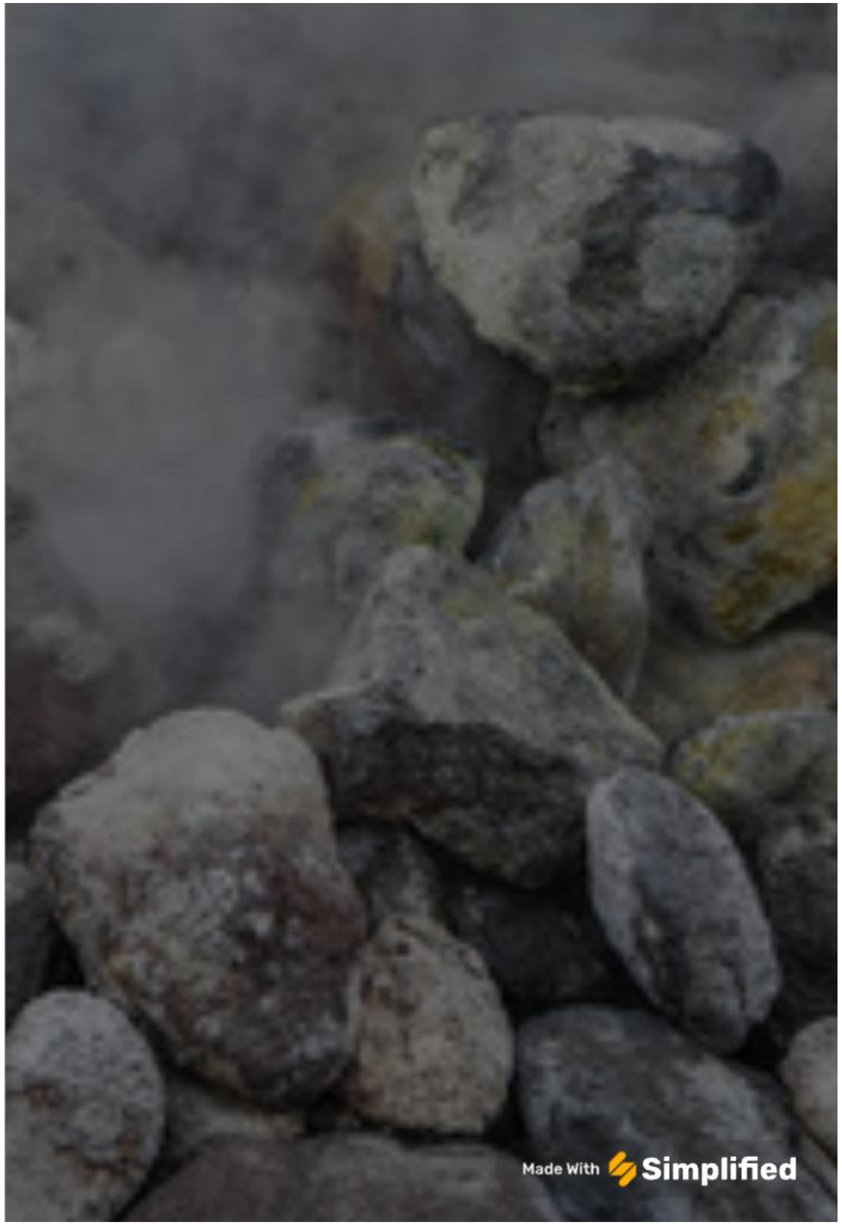


4.Long short-term memory (LSTM)

Long short-term memory networks are recurrent neural networks which are widely applied to model sequential data like time series or natural language. As stated, RNNs suffer from the vanishing gradient problem during backpropagation, and in the case of simple RNNs, their memory barely includes the previous 10 time steps (Bengio et al., 1994). LSTMs, however, can remember long-term dependencies because they have been explicitly designed to overcome this problem (Hochreiter and Schmidhuber, 1997). Besides the hidden state of RNNs, LSTMs have a cell memory (or cell state) to store information and three gates to control the information flow (Hochreiter and Schmidhuber, 1997). The forget gate (Gers et al., 2000) controls which and how much information of the cell memory is forgotten, the input gate controls which inputs are used to update the cell memory, and the output gate controls which elements of the cell memory are used to update the hidden state of the LSTM cell.

5.Data dependency

The data dependency of empirical models is a classical research question (Jakeman and Hornberger, 1993), often focusing on the number of parameters but also concerning the length of available data records. Data scarcity is also an important topic in machine learning in general, especially in deep learning and the focus of recent research (e.g. Gauch et al., 2021). One can therefore expect to find performance differences between both shallow and deep models used in this study. We hence performed experiments to explore the need for training data for each of the model types.



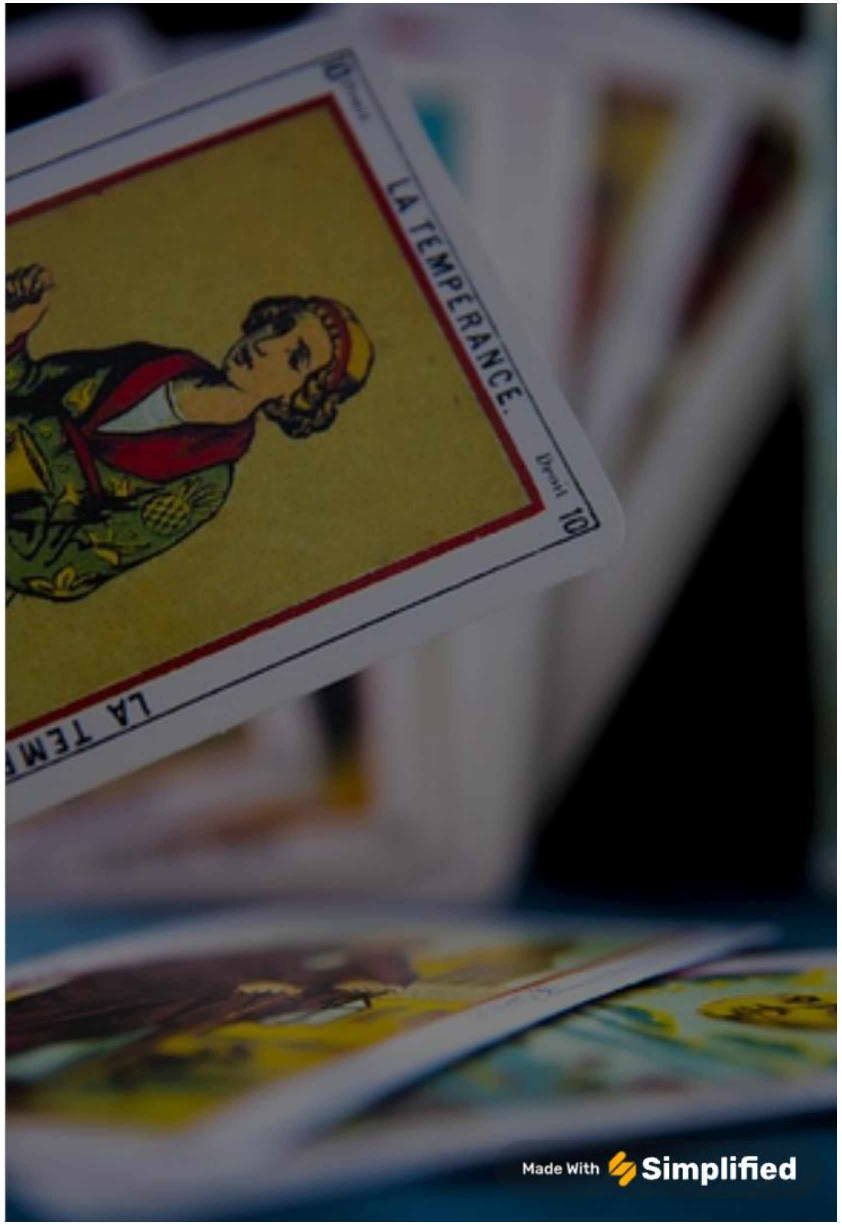
6.Results and discussion

6.1 Sequence-to-value (seq2val) forecasting performance

summarizes and compares the overall seq2val forecasting accuracy of the three model types for all 17 wells. shows the performance when only meteorological inputs are used; the models are additionally provided with GWLt−1 as an input. Because the GWL of the last step has to be known, the latter configuration has only limited value for most applications since only one-step-ahead forecasts are possible in a real-world scenario. However, the inputs of the former configuration are usually available as forecasts themselves for different time horizons.

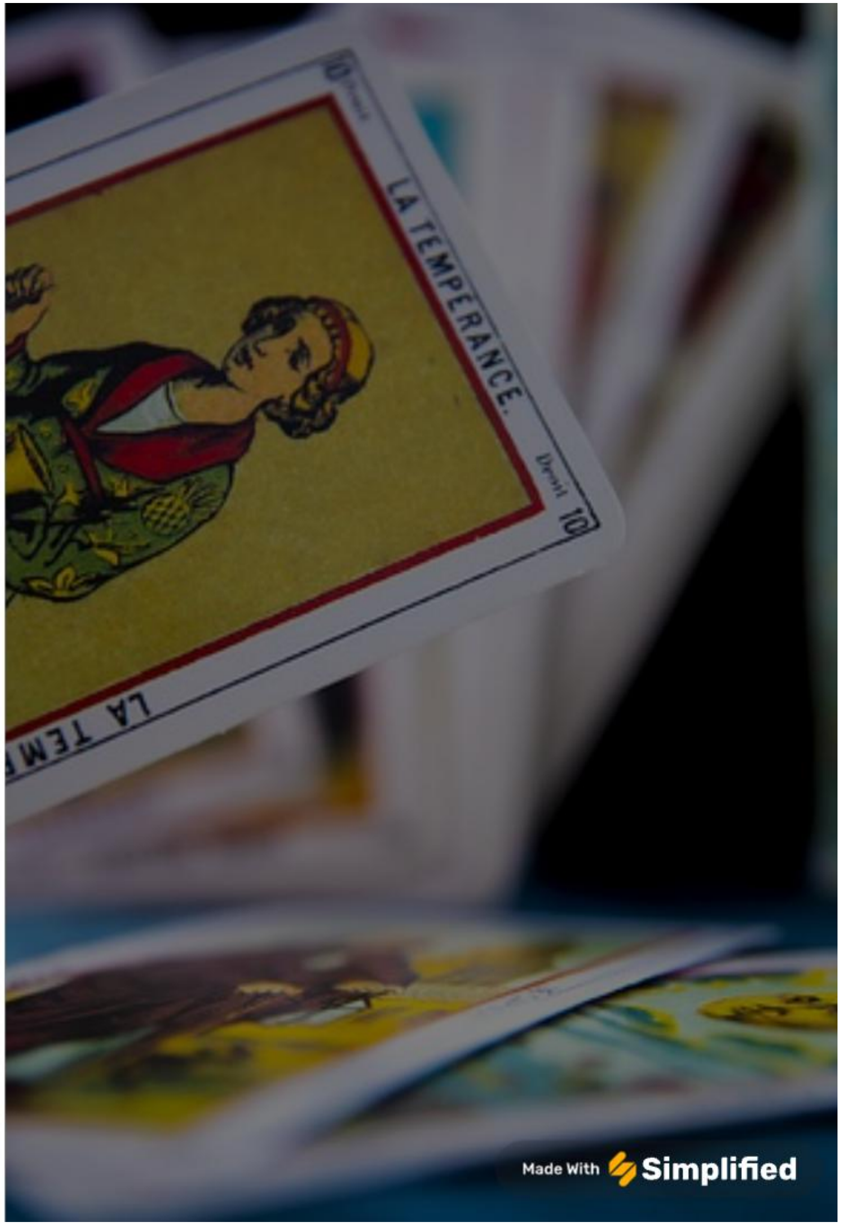
6.2 Sequence-to-sequence (seq2seq) forecasting performance

Sequence-to-sequence forecasting is especially interesting for short- and mid-term forecasts, because the input variables only have to be available until the start of the forecast. summarizes and compares the overall seq2seq forecasting accuracy of the three model types for all 17 wells. shows the performance when only meteorological inputs are used; the models in are additionally provided with GWLt−1 as an input. Similarly to the seq2val forecasts , past GWLs seem to be especially important for LSTM and CNN models, where this additional input variable causes substantial performance improvement.



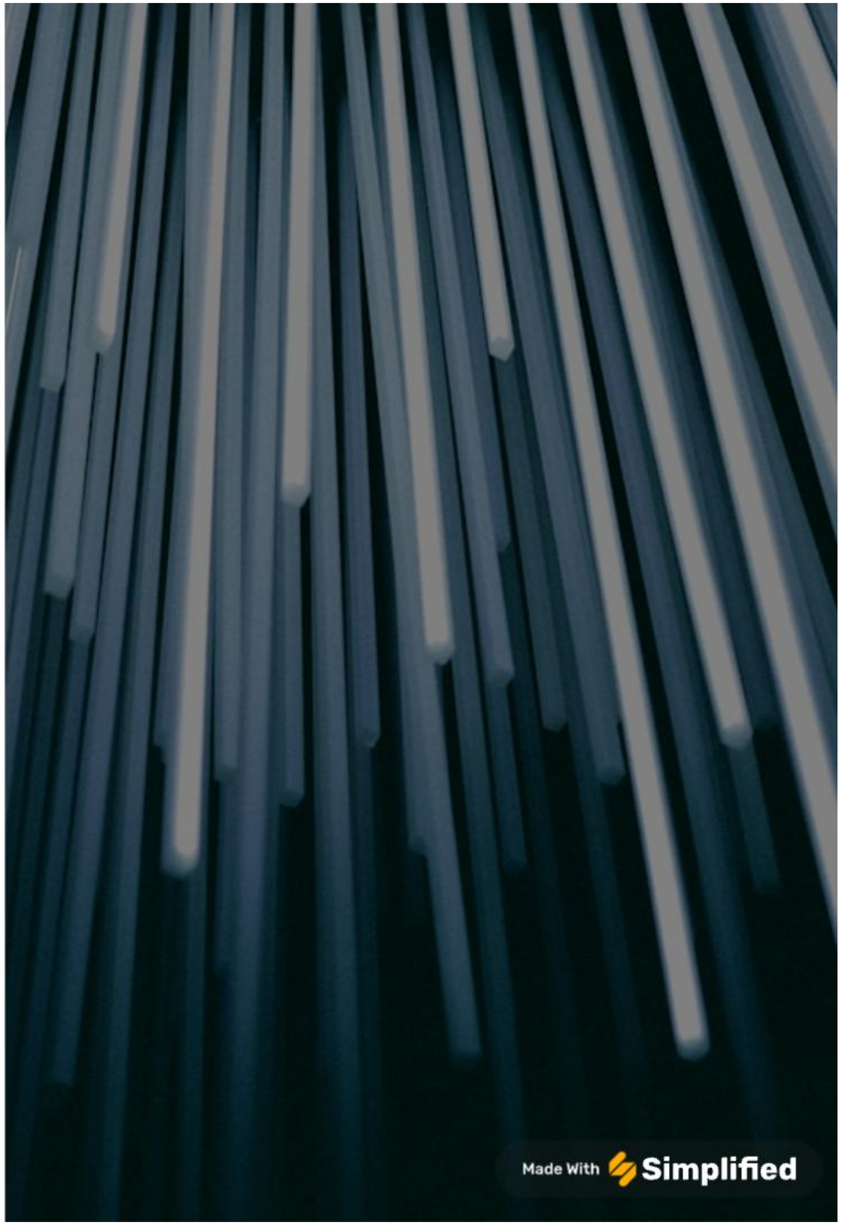
6.3 Hyperparameter optimization and computational aspects

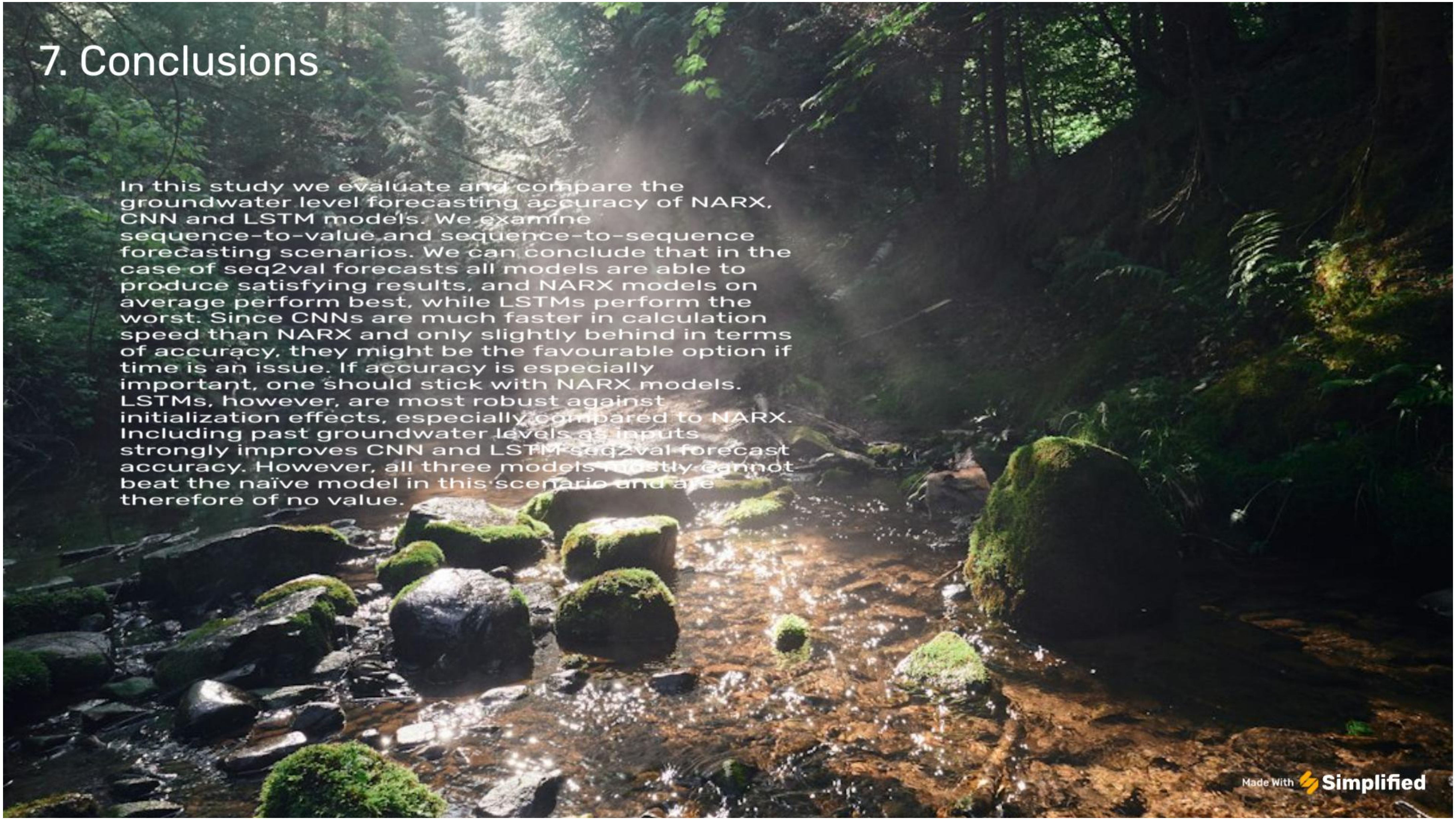
During the HP optimization, depending on the forecasting approach (seq2val/seq2seq) and available inputs (with or without GWLt−1), there were noticeable differences with regard to the number of iterations required and the associated time needed . The best parameter combination, especially for CNN and LSTM networks, was often found in 33 steps or fewer, i.e. after 25 obligatory random exploration steps in only 8 Bayesian steps. Please note that prior to the analysis we chose to at least perform 50 optimization steps, which explains the distribution in the “total iterations” column. In column two (“best iteration”) we can observe similar behaviour of CNNs and LSTMs, while NARX are always somehow different to these two.



6.4 Influence of training data length

In the following section we explore similarities and differences of NARX, LSTMs, and CNNs in terms of the influence of training data length. It is commonly known that data-driven approaches profit from additional data; however, how much data are necessary to build models that are able to perform reasonable calculations still remains an open question. This is because the answer is highly dependent on the application case, data properties (e.g. distribution), and model properties, as model depth can sometimes exponentially decrease the need for training data (Goodfellow et al., 2016). Therefore, this question cannot be entirely answered by a simple analysis like we perform here. Nevertheless, we still want to give an impression of how much data might be approximately needed in the case of groundwater level data in porous aquifers and if the models substantially differ in their need for training data.





7. Conclusions

In this study we evaluate and compare the groundwater level forecasting accuracy of NARX, CNN and LSTM models. We examine sequence-to-value and sequence-to-sequence forecasting scenarios. We can conclude that in the case of seq2val forecasts all models are able to produce satisfying results, and NARX models on average perform best, while LSTMs perform the worst. Since CNNs are much faster in calculation speed than NARX and only slightly behind in terms of accuracy, they might be the favourable option if time is an issue. If accuracy is especially important, one should stick with NARX models. LSTMs, however, are most robust against initialization effects, especially compared to NARX. Including past groundwater levels as inputs strongly improves CNN and LSTM seq2val forecast accuracy. However, all three models mostly cannot beat the naïve model in this scenario and are therefore of no value.