

Short-term predictions of transient shallow groundwater levels at local scale using data-driven models



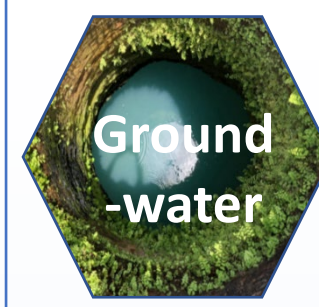
Jiangyue Ju¹, David L. Rudolph¹, Jimmy Lin², Morteza Behbooei²

¹ Department of Earth and Environment Science – University of Waterloo, Waterloo, Ontario, Canada

² David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada



Introduction



- Groundwater (GW) is a critical natural resource for human activities and a vital component of the hydrological cycle.
- Short-term, local-scale groundwater level (GWL) fluctuations are erratic and influenced by several processes^{[1][2]}.



- Fast and accurate prediction of this phenomenon provides critical information for researchers and decision-makers in both the agriculture and environmental sectors. However, it is significantly neglected in favor of long-term GWL studies, primarily due to inherent complexities.



- Physical based models are widely used, but they necessitate comprehensive knowledge of the underlying processes and precise data and parameters.
- Data-driven models are effective and efficient alternatives.
- Versatile, reduced inputs and complexity, independent of system-specific knowledge, quick and cost-effective.

Objectives

- Investigate the feasibility and performance of data-driven models in predicting local, short-term GWLs and temporal recharge events.
- Examine 7 popular data-drive models using time series of hydro-meteorological data. Demonstrates the ease and effectiveness of utilizing these models for short-term GWL prediction.
- Stimulate further discussions and encourage the adoption of data-driven models in hydrological studies, emphasizing their potential benefits and applications.

Study Area

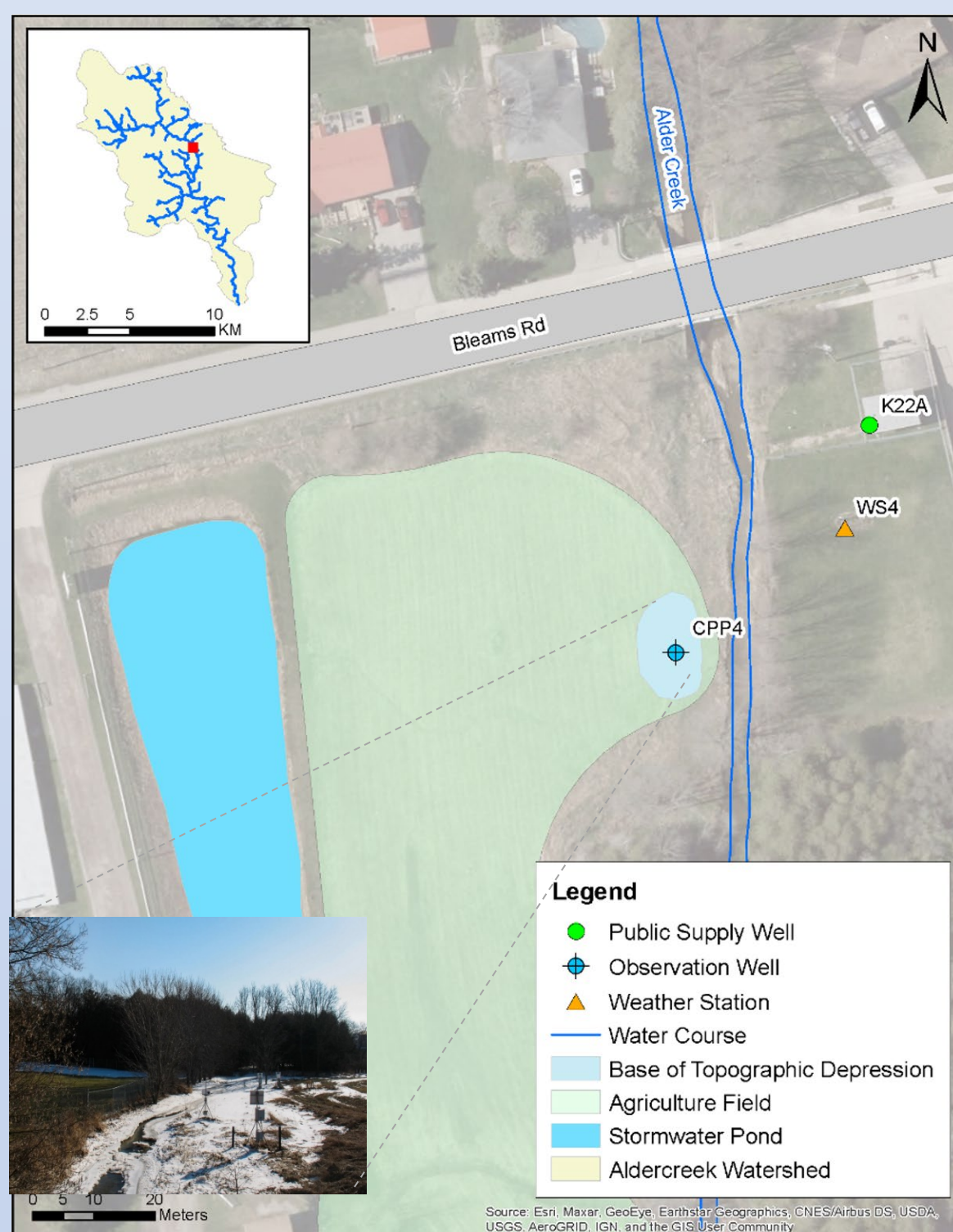


Figure 1: Map of the Mannheim site with instrumentation location and adjacent geographical features. Site photo is credited to A. Wiebe

The Mannheim site is situated within the Aldercreek watershed on the interlobate Waterloo Moraine. It is characterized by its agricultural landscape and features a unique setting that includes a depression base, a dynamic creek, a public supply well, and a stormwater pond. GWL fluctuation prediction at this site is crucial for :

- Identifying recharge events
- Assessing environmental concerns from agricultural contaminants and surface water
- Managing irrigation for local farms.

The site's setting is unique, localized, and complex, making it a challenging task if we use physical hydrological models for prediction.

Study Approach

Feature selection and data Description

Daily Time Series (2 y)

- Historical GWL
- Precipitation
- Evapotranspiration
- Temperature

Pre-Processing

- Filling:
 - Non-continuous
 - Adjacent station (IDW)

Smoothing:

- Wavelet-transform

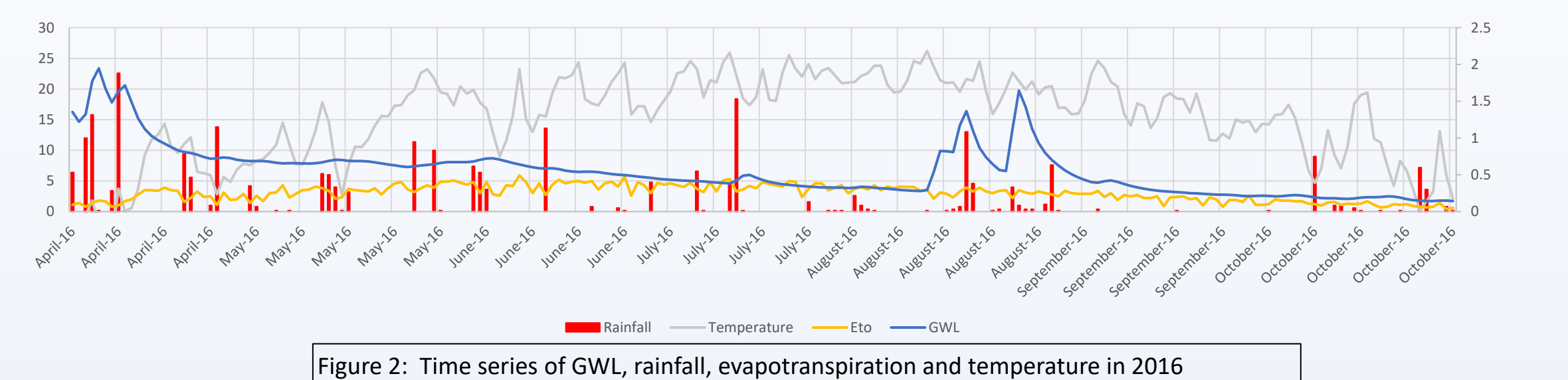


Figure 2: Time series of GWL, rainfall, evapotranspiration and temperature in 2016

Model selection and structure

7 data-driven models are developed in Python to provide a good variety of current popular data-driven machine learning algorithm.

- Linear Regression (LR)
- Autoregressive Integrated Moving Average with exogenous variables (ARIMAX)
- Feedforward Neural Network (FFNN)
- Recurrent Neural Network (RNN)
- Support Vector Regression (SVR)
- Random Forest Regressor(RFR)
- K-nearest Neighbor (KNN)

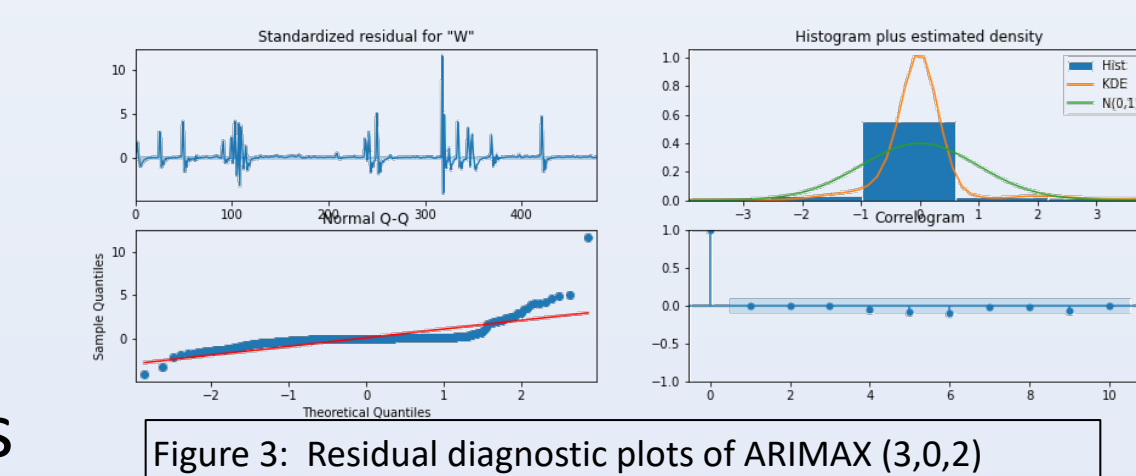


Figure 3: Residual diagnostic plots of ARIMAX (3,0,2)

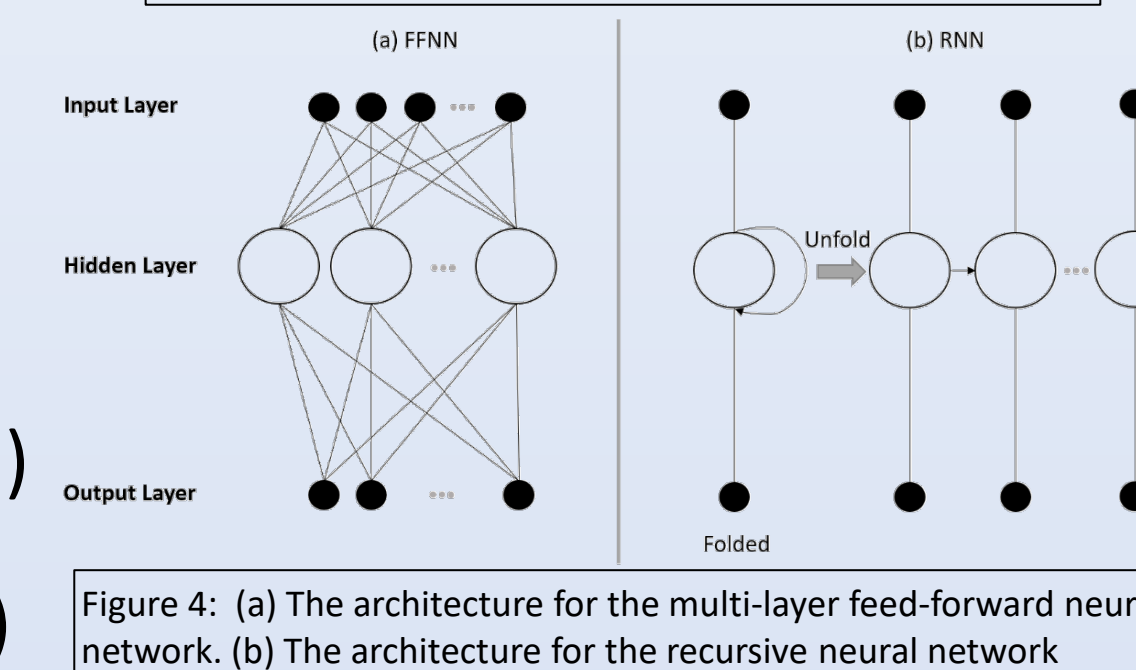


Figure 4: (a) The architecture for the multi-layer feed-forward neural network. (b) The architecture for the recurrent neural network

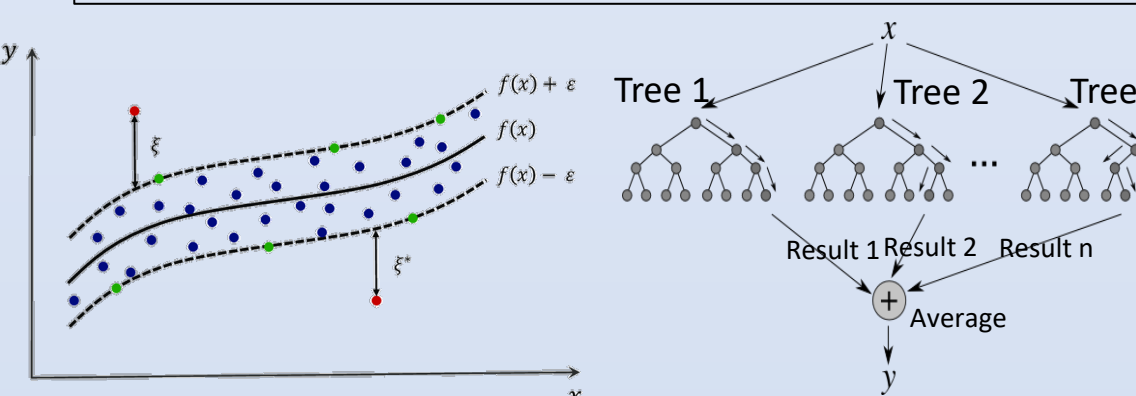


Figure 5: Sketch of a nonlinear SVR

Figure 6: Structure of a RFR

1-day sliding window rolling forecast is used across all algorithms with a 3-day window size based on hydrological considerations, also validated through autocorrelation and cross-correlation analyses.

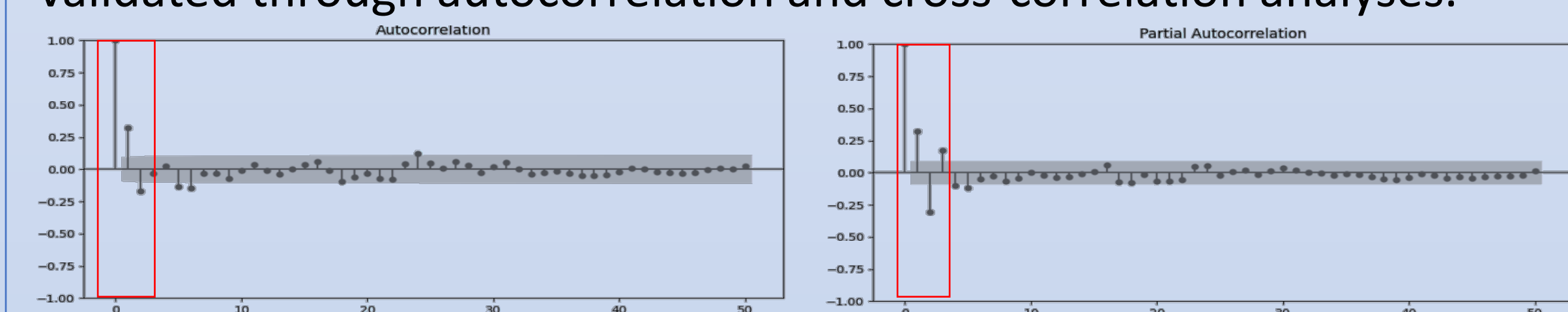


Figure 7: Correlation functions for GWL

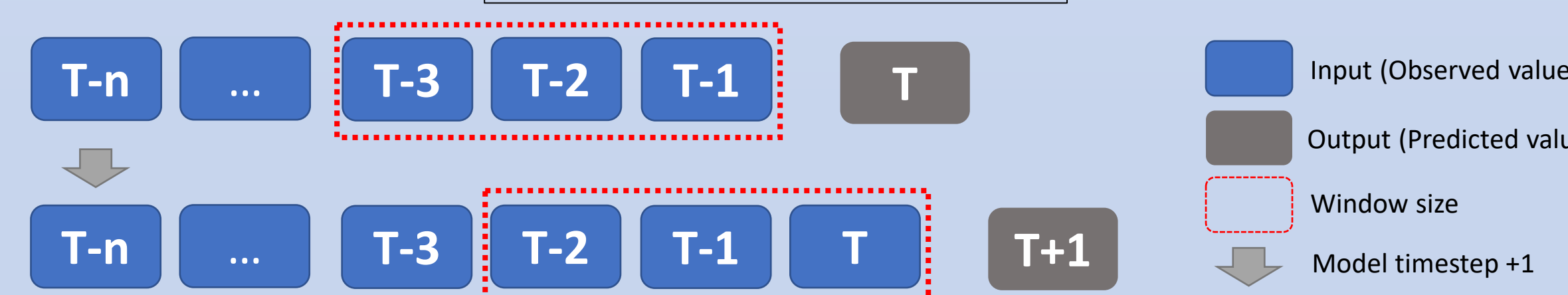


Figure 8: Diagram of sliding window rolling forecast

Training, testing and evaluation criteria

Models were trained and validated using an 80:20 train-test split.

The performance of the models was evaluated using:

Root mean square error (RMSE), correlation coefficient (R2), and Mean absolute error (MAE).

Results

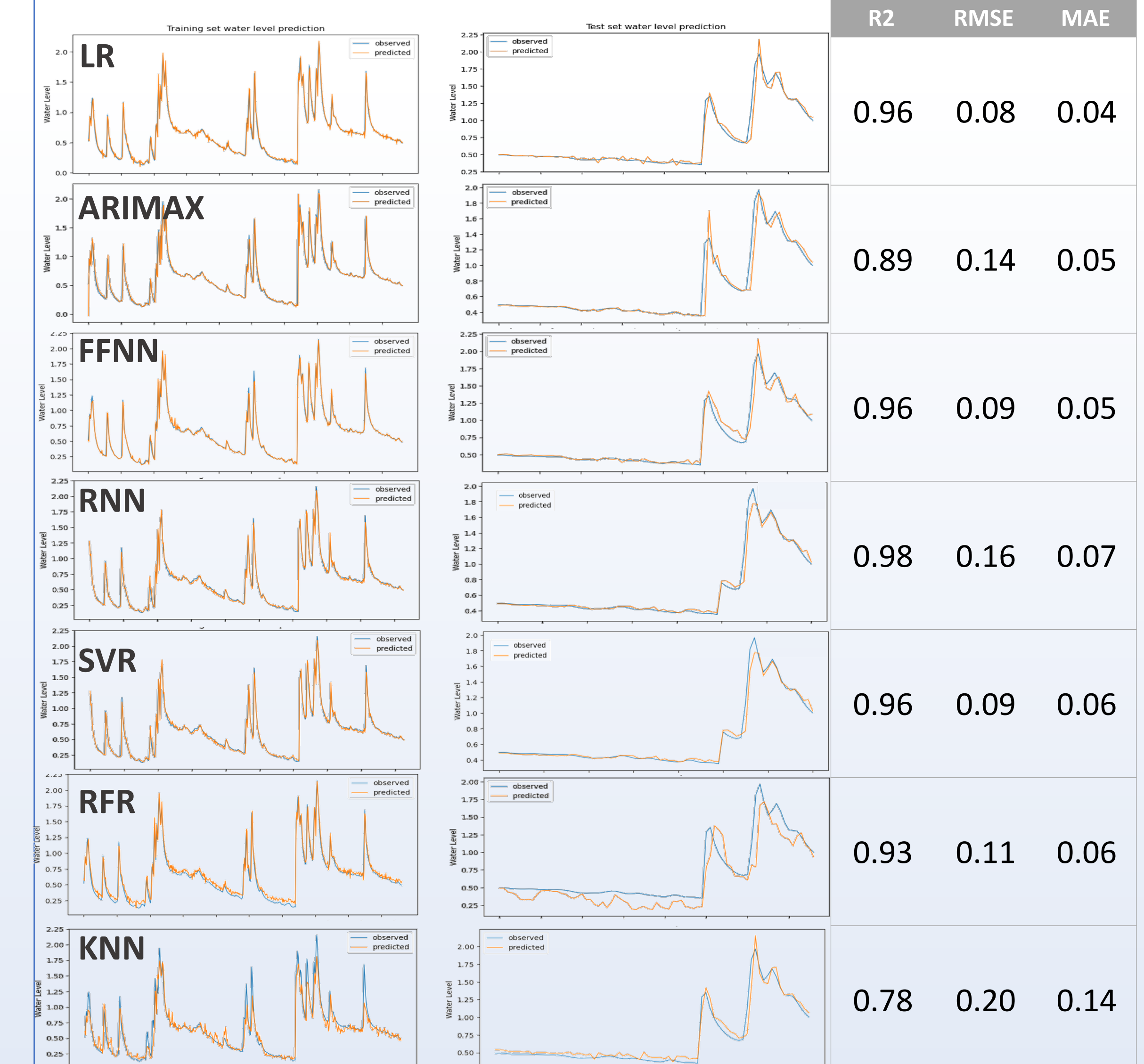


Figure 9: Training and testing results of each data-driven model

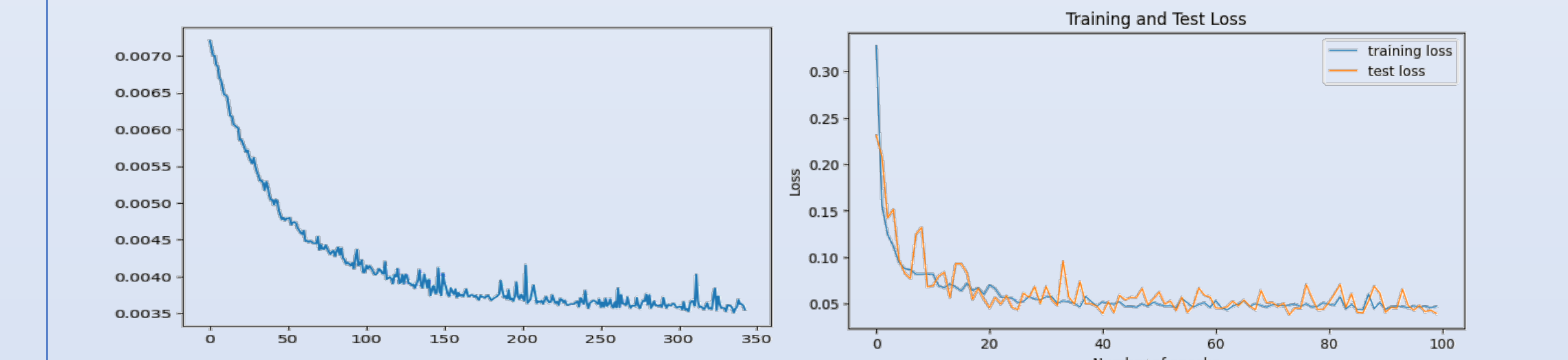


Figure 10: Loss function of neural network models

- Statistical algorithms such as ARIMAX and classification algorithms like KNN struggles in achieving accurate predictions.
- RNN, specifically designed for such problems, has the potential to excel but is prone to overfitting and requires careful tuning.
- Special attention should be given to the issue of peak shifting.

Conclusion and Future Steps

- Data-driven models can effectively predict short-term local GWL fluctuations, even with limited hydro-meteorological data.
- All 7 models yielded acceptable predictions using just 2 years of training data, each with distinct strengths and weaknesses.
- Evaluation should not solely rely on metrics.
- Among 7 models, RNN demonstrates the highest performance.

This approach needs to be applied to a more common setting to demonstrate universal application. A comparison and integration of data-driven models with physical models is suggested.

References and Acknowledgements

- [1] Cherry, J. A., & Freeze, R. A. (1979). Groundwater: Englewood Cliffs, NJ: Prentice-Hall.
- [2] Gribovski, Z., Szilágyi, J., & Kalicz, P. (2010). Diurnal fluctuations in shallow groundwater levels and streamflow rates and their interpretation – A review. Journal of Hydrology, 385(1), 371-383.
- The authors would like to express special thanks for contributions from Andrew Wiebe and Sheldon Liu