

Monthly Groundwater Level Simulation in Shallow Dutch Aquifers: Transfer Function-Noise Modelling versus Machine Learning

Eshagh Bandani¹, Rahleh Ahmadi²

¹ Contracts Affairs Expert, Sistan and Baluchestan Regional Water Company, Zahedan, Iran. ()
eshaghibandani@gmail.com)

² Head of the Water Resources Balancing Group, Sistan and Baluchestan Regional Water Company, Zahedan, Iran.

Received: 11 May 2025

Accepted: 18 June 2025

Available online: 25 June 2025

Abstract: Groundwater level time series are increasingly simulated with both physically motivated transfer function-noise (TFN) models and purely data-driven machine learning (ML) models, but the two families are rarely compared under an identical, forcing-only protocol. This study compared a TFN model (Pastas, gamma impulse response, linear recharge, AR(1) noise) with random forest (RF), extreme gradient boosting (XGBoost), support vector regression (SVR) and a long short-term memory (LSTM) network for simulating monthly mean hydraulic heads in three observation wells in the Netherlands (B58C0698 near Heibloem; B32C0609 and B32C0639 near De Bilt), with records of 24–37 years. All models were driven only by precipitation and reference evaporation, without lagged heads, and were trained on the first 70–80% of each record and tested on the remaining 79–95 months. Heads showed a regular winter–spring maximum and late-summer minimum and were most strongly correlated with the 6-month cumulative precipitation surplus ($r = 0.74–0.93$). No significant trend was found in annual heads or annual precipitation surplus. In the test periods, all models performed well at B58C0698 (NSE = 0.90–0.93) and B32C0639 (NSE = 0.79–0.87), with differences between models that were not statistically significant (Diebold–Mariano $p > 0.27$). At B32C0609, whose test period coincided with the start of daily logger measurements, all models underestimated heads by 8–12 cm and NSE fell to 0.36–0.58. The tree ensembles were the least affected but not significantly better than TFN ($p \geq 0.05$). Averaged over the three wells, SVR (NSE = 0.76) and RF (0.76) ranked slightly above XGBoost (0.75), LSTM (0.71) and TFN (0.70). Permutation importance attributed 54–91% of the ML skill to the 6-month precipitation surplus, and the TFN step responses reached 95% of their final value after 323–441 days. With forcing-only inputs and a few hundred monthly observations, a parsimonious TFN model is as accurate as the ML models and easier to interpret, whereas the ML models add value mainly where the response is not captured by a linear recharge assumption.

Keywords: Groundwater level, Transfer function-noise model, Pastas, LSTM, Random forest, Precipitation surplus

1. Introduction

Groundwater provides a large share of the world's drinking and irrigation water, and hydraulic heads are the most widely monitored indicator of groundwater storage. Long head records are used to assess the effects of climate variability, abstraction and land-use change, to detect drought, and to design monitoring networks (Taylor et al., 2013; Bloomfield and Marchant, 2013). Where data or resources for a distributed groundwater flow model are lacking, time series models that relate observed heads directly to their driving stresses are an efficient alternative (Bakker and Schaars, 2019).

Two families of time series models dominate. The first is the transfer function-noise (TFN) approach, in which the head is the convolution of a stress, usually the precipitation surplus, with a parametric impulse response function, plus a noise model. Von Asmuth et al. (2002) formulated the method in continuous time with predefined response functions, and the open-source package Pastas (Collenteur et al., 2019) has made it widely accessible. TFN models have few parameters, and their response functions can be interpreted in terms of aquifer properties and recharge (Oberghell et al., 2019). Non-linear recharge formulations extend them to conditions in which the soil-moisture state matters (Peterson and Western, 2014).

The second family comprises machine learning (ML) models such as artificial neural networks, support vector regression (SVR), tree ensembles and recurrent networks. Early studies showed that neural networks can forecast heads in shallow aquifers (Nayak et al., 2006). Later reviews summarised a rapid growth of ML applications for groundwater level modelling (Rajaei et al., 2019). Sahoo et al. (2017) used ML to model groundwater level changes in agricultural regions of the United States, and Wunsch et al. (2021) showed that long short-term memory (LSTM) and convolutional networks can forecast heads from meteorological inputs alone. Many ML studies, however, use lagged observed heads as inputs. This makes the task close to one-step persistence and makes it hard to compare the results with simulation models driven only by meteorological stresses.

The 2022 Groundwater Time Series Modelling Challenge compared many data-driven and TFN models on the same wells and forcing data and found that no single approach performed best everywhere (Collenteur et al., 2024). Such comparisons remain scarce for monthly data, which is the resolution of many national monitoring programmes and water-resources assessments. Monthly data also yield only a few hundred samples per well, a small training set for data-hungry ML models.

The objective of this study was therefore to compare a TFN model with four ML models (RF, XGBoost, SVR and LSTM) for simulating monthly groundwater levels in three Dutch observation wells, using forcing-only inputs and an independent test period. The specific aims were to (i) characterise the seasonal dynamics, forcing relationships and trends of the heads; (ii) compare the accuracy and significance of differences between the models; and (iii) interpret the models through TFN response functions and ML permutation importance. All data are openly available, so the analysis can be reproduced.

2. Materials and Methods

2.1. Study Sites and Data

Three observation wells from the Dutch national subsurface database (DINO, TNO) were used, as distributed with the example data of the Pastas package (Collenteur et al., 2019) (Figure 1; Table 1). Well B58C0698 lies near Heibloem in the province of Limburg, in the southern Netherlands. Its heads were modelled with daily precipitation from the Heibloem rain gauge and daily reference evaporation from the Maastricht weather station. Wells B32C0609 and B32C0639 lie near De Bilt in the province of Utrecht and were modelled with daily precipitation and Makkink reference evaporation from KNMI station De Bilt (station 260). Heads are expressed in metres relative to the Dutch ordnance datum (NAP). Most heads were measured manually about twice a month, but B32C0609 was measured almost daily from September 2010 onward.



Figure 1. Approximate location of the observation wells and meteorological stations in the Netherlands

Monthly mean heads were computed from all measurements in each calendar month, and monthly precipitation (P) and evaporation (E) were computed as sums of the daily values. To allow a warm-up period for the ML inputs, the analysis started in January 1982 at the De Bilt wells and in November 1985 at B58C0698, and it ended with the last observation (Table 1). Each record was split chronologically: the training periods ended in December 2007 (B58C0698), December 2009 (B32C0609) and December 1998 (B32C0639), and the remaining months (79–95 per well) were used for testing only.

Table 1. Characteristics of the observation wells and data (monthly analysis period)

Well	Location	Meteorological data	Period	Head obs.	Months	Train/test months	Mean head (m)	Range (m)	SD (m)	P (mm yr ⁻¹)	E (mm yr ⁻¹)
B58C0698	Heibloem, Limburg	Heibloem (P) / Maastricht (E)	1985-11 – 2015-06	644	341	253/88	27.90	26.71–28.89	0.42	752	590
B32C0609	De Bilt, Utrecht	De Bilt (KNMI 260)	1982-01 – 2018-02	3200	411	316/95	1.22	0.51–2.12	0.30	837	564
B32C0639	De Bilt, Utrecht	De Bilt (KNMI 260)	1982-01 – 2005-10	532	280	201/79	1.36	0.95–1.96	0.20	824	558

Heads in m relative to NAP. P = precipitation; E = reference evaporation (Makkink). Head obs. = number of individual measurements.

2.2. Transfer Function-Noise Model

The TFN model simulates the head h as the response to recharge R plus a constant and a noise term (von Asmuth et al., 2002; Collenteur et al., 2019):

$$h(t) = \int_{-\infty}^t R(\tau) \theta(t - \tau) d\tau + d + n(t) \quad (1)$$

where θ is the impulse response function, d is the base level and n is the residual series. Recharge was approximated by the linear model $R = P + fE$, where f is an evaporation factor. The impulse response was a scaled gamma distribution:

$$\theta(t) = A \frac{t^{n-1}}{a^n \Gamma(n)} e^{-t/a} \quad (2)$$

where A is the gain (the steady-state head change caused by a unit constant recharge), and n and a are shape and scale parameters. The mean response time is $n \cdot a$. The residuals were described by an AR(1) noise model with characteristic time α :

$$n(t_i) = e^{-(t_i - t_{i-1})/\alpha} n(t_{i-1}) + v(t_i) \quad (3)$$

where v is white noise. The six parameters (A , n , a , f , d , α) were estimated at daily resolution by least squares on the noise innovations over the training period, with a 10-year warm-up. The calibrated model was then run over the whole record without noise, and the simulated daily heads were averaged over the observation dates of each month so that they matched the monthly observed means. The Pastas package (version 2.0) was used.

2.3. Machine Learning Models

The ML models were trained at monthly resolution in simulation mode, i.e. without lagged observed heads. The inputs (Table 2) were monthly P , E and precipitation surplus $P - E$, the 3-month sums of P and E , the cumulative surplus over 3, 6, 12 and 24 months, and a sine-cosine encoding of the month. The cumulative surplus was defined as:

$$S_k(t) = \sum_{j=0}^{k-1} [P(t-j) - E(t-j)], \quad k \in \{3, 6, 12, 24\} \quad (4)$$

Table 2. Input variables of the machine learning models

Input	Definition
P, E	Monthly precipitation and reference evaporation (mm)
P – E	Monthly precipitation surplus (mm)
ΣP 3 mo, ΣE 3 mo	Sums of P and E over the current and two previous months (mm)
$\Sigma P - E$ k mo	Cumulative surplus S_k over $k = 3, 6, 12, 24$ months (mm)
$\sin(m)$, $\cos(m)$	Seasonal encoding of the calendar month m

Random forest (RF) averages 500 regression trees grown on bootstrap samples with random subsets of inputs (Breiman, 2001). **XGBoost** builds an additive ensemble of shallow trees by gradient boosting with second-order optimisation and regularisation (Chen and Guestrin, 2016). **SVR** with a radial basis function kernel fits a flat function within an ε -insensitive tube (Vapnik, 1995). The hyperparameters of these three models were tuned by grid search with five-fold forward-chaining cross-validation on the training period (Table 3), and all inputs were standardised.

The **LSTM** network (Hochreiter and Schmidhuber, 1997) received sequences of the 24 preceding months of standardised P , E and the seasonal terms. Its memory cell is updated through input (i), forget (f) and output (o) gates:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c[h_{t-1}, x_t] + b_c) \quad (5)$$

$$h_t = o_t \odot \tanh(c_t) \quad (6)$$

where each gate is a sigmoid function of the previous hidden state h_{t-1} and the current input x_t . The network had one LSTM layer with 16 units, dropout of 0.2 and a linear output neuron. It was trained with the Adam optimiser (learning rate 0.002, batch size 16) and early stopping (patience 30 epochs) on the last 20% of the training period. The prediction was the mean of three networks initialised with different random seeds. The ML models were implemented with scikit-learn (Pedregosa et al., 2011), xgboost and Keras.

Table 3. Hyperparameter search spaces and selected values (in the order listed; wells identified by their last four digits)

Model	Search space	Selected
RF	max depth $\in \{4, 8, \text{none}\}$; min samples leaf $\in \{2, 5\}$; max features $\in \{0.33, 0.66, 1.0\}$	0698: None/2/0.66; 0609: None/2/1.0; 0639: 8/2/0.33
XGBoost	trees $\in \{200, 500\}$; learning rate $\in \{0.02, 0.05\}$; max depth $\in \{2, 3, 5\}$	0698: 200/0.02/3; 0609: 200/0.05/2; 0639: 200/0.02/2
SVR	$C \in \{0.1, 1, 10\}$; $\gamma \in \{0.01, 0.05, 0.1\}$; $\varepsilon \in \{0.01, 0.05\}$	0698: 1/0.01/0.01; 0609: 10/0.01/0.01; 0639: 1/0.01/0.05
LSTM	fixed: 16 units, dropout 0.2, sequence 24 months, 3-seed ensemble	early stopping

2.4. Trend Analysis and Evaluation

Trends in annual mean heads (years with at least 10 monthly values) and annual precipitation surplus were tested with the Mann–Kendall test with the Hamed and Rao (1998) correction for serial correlation. For one series, where the corrected variance could not be computed, the original test was used (Mann, 1945; Kendall, 1975). Slopes were estimated with the Sen (1968) method. Model performance was measured with the Nash–Sutcliffe efficiency (NSE; Nash and Sutcliffe, 1970), root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), mean bias and the Kling–Gupta efficiency (KGE; Gupta et al., 2009):

$$NSE = 1 - \frac{\sum_i (O_i - S_i)^2}{\sum_i (O_i - \bar{O})^2} \quad (7)$$

$$KGE = 1 - \sqrt{(r - 1)^2 + \left(\frac{\sigma_S}{\sigma_O} - 1\right)^2 + \left(\frac{\mu_S - \mu_O}{\sigma_O}\right)^2} \quad (8)$$

where O and S are observed and simulated monthly heads, r is their correlation, and σ and μ are standard deviations and means. Because heads refer to an arbitrary datum, the bias component of KGE was written as the mean error divided by σ_O rather than as the ratio of the means. This makes the score independent of the datum. Differences in squared errors between each ML model and TFN were tested with the Diebold–Mariano test (Diebold and Mariano, 1995) using a variance estimate that allows lag-1 autocorrelation. Input relevance in the ML models was measured by permutation importance on the test period (30 repetitions), averaged over RF and XGBoost.

3. Results

3.1. Groundwater Level Dynamics

The heads varied within about 2.2 m at B58C0698 and within about 1.6 m and 1.0 m at the two De Bilt wells (Table 1; Figure 2). All three wells showed the same seasonal cycle (Figure 3). Heads peaked in February–March, after the winter precipitation surplus, and reached their minimum in August–September, when evaporation exceeds precipitation. The mean seasonal amplitude was 0.95 m at B58C0698, compared with 0.42 m and 0.33 m at B32C0609 and B32C0639. The dry years 1996 and 2003 produced the lowest heads in the De Bilt records, and the wet winters of 1987–1988 and 1998–1999 produced the highest heads (Figure 2).

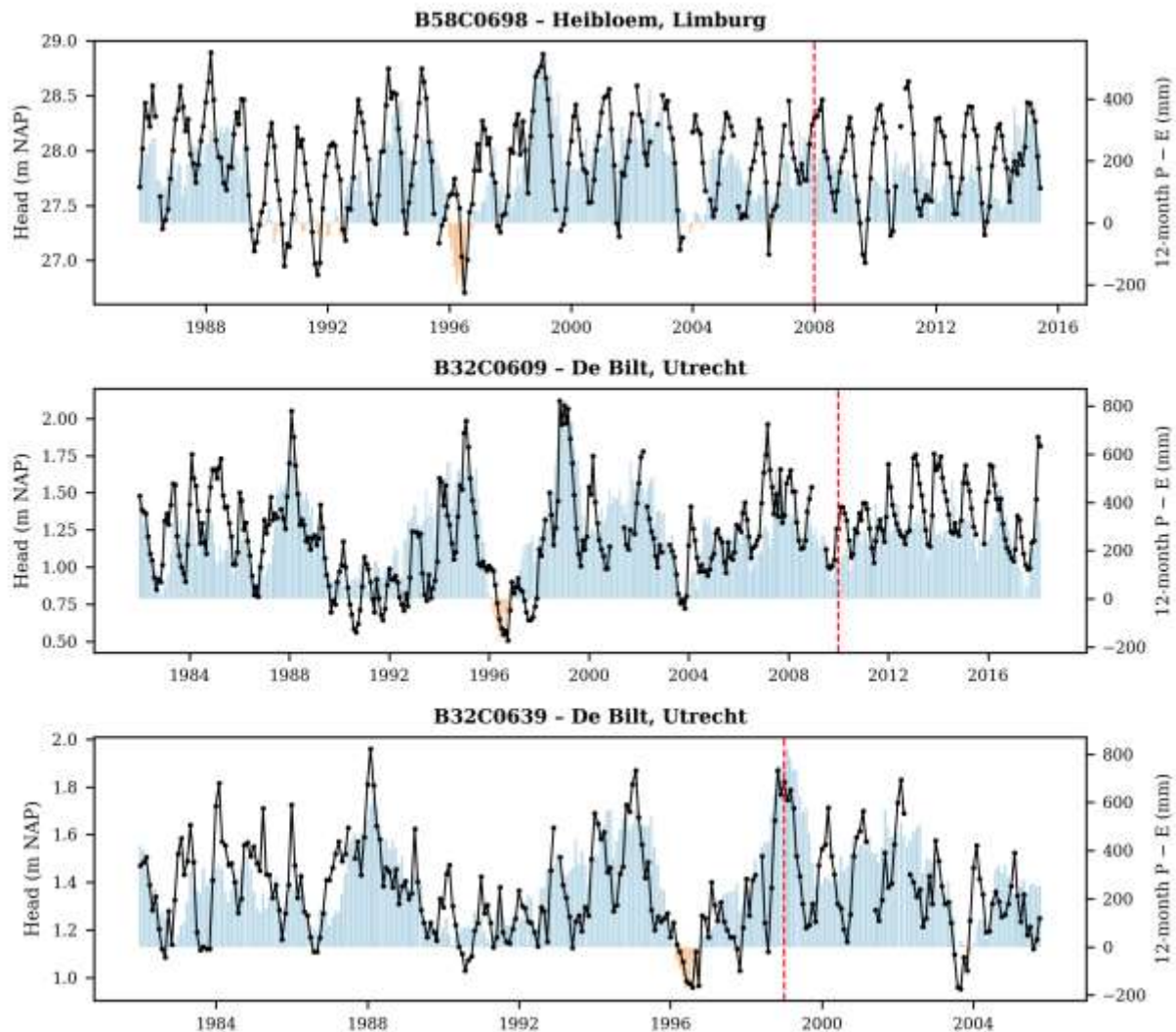


Figure 2. Monthly mean heads (black) and 12-month cumulative precipitation surplus (bars) at the three wells; the red dashed line marks the end of the training period

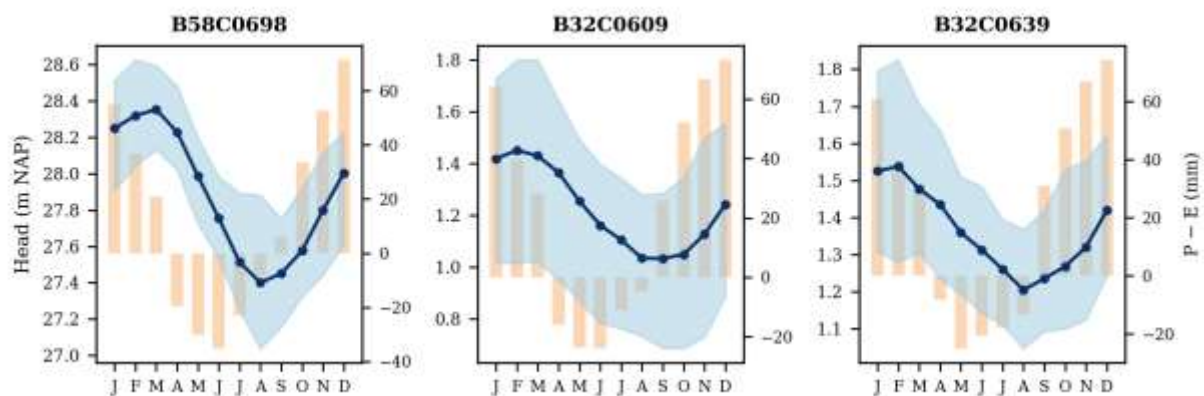


Figure 3. Seasonal regime of monthly heads (line: mean; shading: 10th–90th percentiles) and mean monthly precipitation surplus (bars)

Heads were most strongly correlated with the 6-month cumulative surplus ($r = 0.93, 0.74$ and 0.82 for B58C0698, B32C0609 and B32C0639). Correlations with the 3-month surplus (0.52 – 0.73) and the 12-month surplus (0.49 – 0.64) were lower. The trend tests (Table

4) found no significant change in annual mean heads ($|Z| \leq 1.61, p \geq 0.11$) or annual precipitation surplus ($p \geq 0.51$) at any well. The positive Sen slope at B32C0609 ($+5.7$ cm per decade) was therefore not significant, although it reflects higher heads after 2010.

Table 4. Trends in annual mean heads and annual precipitation surplus (P – E)

Well	Years	Z (head)	p	Sen (cm decade ⁻¹)	Z (P – E)	p	Sen (mm yr ⁻¹)
B58C0698	27	-0.15	0.88	-0.9	0.61	0.54	+1.7
B32C0609	32	1.61	0.11	+5.7	0.66	0.51	+2.5
B32C0639	24	-0.93	0.35	-1.9	0.62	0.54	+3.0

Modified Mann–Kendall test (Hamed and Rao, 1998), except P – E at B32C0609 (original test).

Table 5 and Figures 4 and 5 summarise the performance of the models. At B58C0698, all models reproduced the test-period heads closely (NSE = 0.90–0.93; RMSE = 9.7–12.0 cm). SVR was the best (NSE = 0.93, KGE = 0.96) and TFN was second (0.92). At B32C0639 the LSTM had the highest skill (NSE = 0.87, RMSE = 7.0 cm), followed by TFN (0.82) and SVR (0.82). The tree ensembles reached 0.79–0.79.

3.2. Model Performance

Table 5. Performance of the models in the training (NSE) and test periods

Well	Model	Train NSE	Test NSE	Test KGE	Test R ²	Test RMSE (cm)	Test MAE (cm)	Test bias (cm)
B58C0698	Pastas-TFN	0.94	0.92	0.92	0.92	10.9	9.0	-2.5
	RF	0.98	0.91	0.94	0.91	11.4	9.0	-1.6
	XGBoost	0.96	0.90	0.94	0.90	12.0	9.3	-1.5
	SVR	0.93	0.93	0.96	0.94	9.7	7.7	-0.8
	LSTM	0.90	0.91	0.94	0.91	11.1	9.0	+1.6
B32C0609	Pastas-TFN	0.71	0.37	0.38	0.75	16.3	13.3	-12.4
	RF	0.95	0.57	0.57	0.74	13.4	10.7	-8.2
	XGBoost	0.89	0.55	0.56	0.72	13.8	10.7	-8.2
	SVR	0.80	0.52	0.45	0.81	14.2	11.9	-11.0
	LSTM	0.79	0.36	0.37	0.73	16.5	13.4	-12.4
B32C0639	Pastas-TFN	0.71	0.82	0.78	0.86	8.2	6.0	+3.8
	RF	0.93	0.79	0.73	0.82	8.9	7.1	+2.3
	XGBoost	0.85	0.79	0.74	0.82	8.8	7.3	+2.2
	SVR	0.78	0.82	0.74	0.87	8.3	6.4	+4.2
	LSTM	0.79	0.87	0.84	0.88	7.0	5.7	+2.3

Bold: highest test NSE per well. Bias = mean(simulated – observed).

The test period of B32C0609 was more difficult. All models underestimated the heads, with mean biases of -8.2 cm (RF) to -12.4 cm (LSTM), and NSE fell to 0.36–0.57. The correlation between simulated and observed heads remained reasonable ($R^2 = 0.72$ –0.81), and the TFN model had one of the higher R^2 values (0.76). The loss of skill therefore came mainly from a systematic

offset rather than from poor timing. This offset appeared after September 2010, when daily measurements replaced the fortnightly manual readings, and was largest in 2013–2015 (Figure 4). RF and XGBoost were the least biased and achieved the highest NSE at this well (0.58 and 0.55).

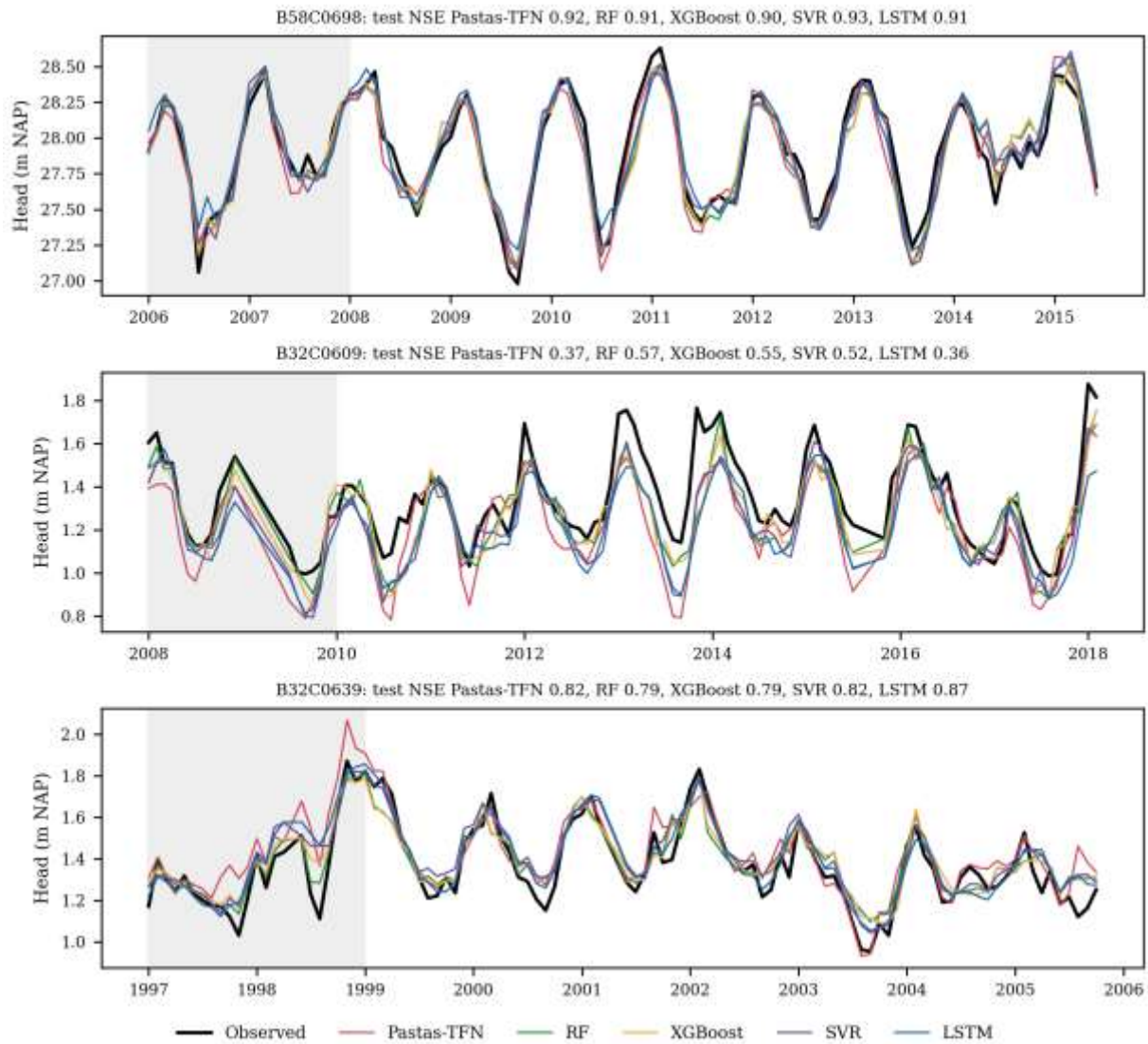


Figure 4. Observed and simulated monthly heads in the test periods (white background) and the last two years of the training periods (grey)

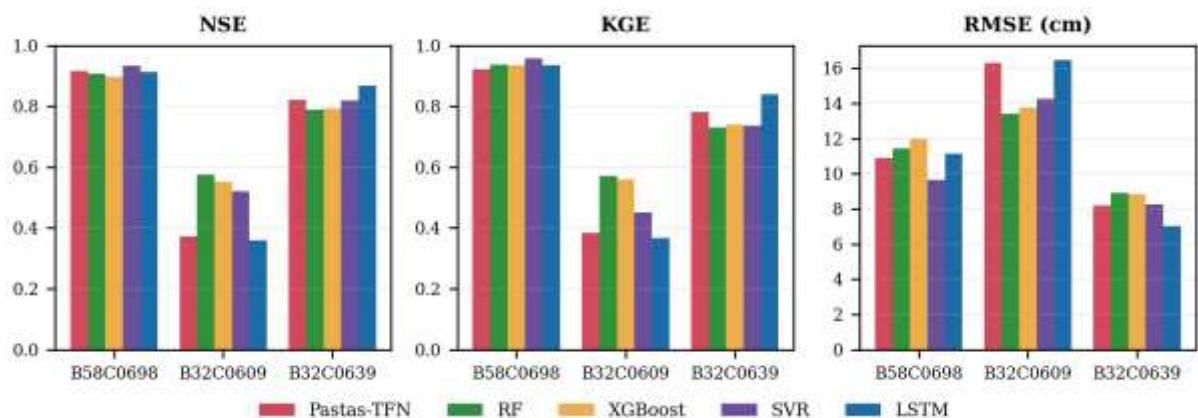


Figure 5. Test-period NSE, KGE and RMSE of the five models at the three wells

Averaged over the three wells, SVR (mean test NSE = 0.76) and RF (0.76) ranked slightly above XGBoost (0.75), LSTM (0.71) and TFN (0.70). This ranking depends almost entirely on B32C0609. None of the ML

models was significantly more accurate than TFN at the 5% level (Table 6). At B58C0698 and B32C0639 all p -values exceeded 0.27. At B32C0609 the tree ensembles

and SVR were marginally better than TFN ($p = 0.05$ – 0.09), whereas the LSTM was not ($p = 0.92$).

The train–test NSE gaps showed that RF and XGBoost fitted the training data much more closely than the other models (training NSE up to 0.95). At B58C0698 and B32C0639 this higher training fit did not translate into better test performance. TFN, SVR and LSTM had similar training and test scores, and at B32C0639 their test NSE was even higher than their training NSE. This reflects the smaller head variability in that test period.

Table 6. Diebold–Mariano tests of squared test errors: machine learning models versus TFN

Well	Model	DM	p	n
B58C0698	RF	0.47	0.642	88
B58C0698	XGBoost	0.87	0.386	88
B58C0698	SVR	-1.09	0.277	88
B58C0698	LSTM	0.21	0.837	88
B32C0609	RF	-1.96	0.053	95
B32C0609	XGBoost	-1.71	0.091	95
B32C0609	SVR	-1.84	0.069	95
B32C0609	LSTM	0.10	0.918	95
B32C0639	RF	0.66	0.512	79
B32C0639	XGBoost	0.61	0.545	79
B32C0639	SVR	0.07	0.941	79
B32C0639	LSTM	-1.03	0.306	79

Negative DM: smaller squared errors than TFN.

3.3. Model Interpretation

The calibrated TFN parameters (Table 7) show physically consistent responses. The gains were largest at B58C0698 ($A = 0.68$ m per mm d⁻¹), where the seasonal amplitude is largest, and smallest at B32C0639 (0.30). The evaporation factors ranged from -0.52 to -1.23. The step responses (Figure 6) reached half of their final value after 70–115 days and 95% after 323–441 days. This agrees with the dominance of the 6-month

surplus in the correlation analysis. The TFN models explained 71–93% of the head variance in the training periods.

Table 7. Calibrated TFN parameters and response characteristics

Well	A (m per mm d ⁻¹)	n	a (d)	f	Mean response time (d)	t50 (d)	t95 (d)	EVP train (%)
B58C0698	0.678	1.02	145	-1.23	148	104	441	93.2
B32C0609	0.562	1.31	115	-0.60	151	115	412	70.6
B32C0639	0.298	0.91	114	-0.52	104	70	323	70.8

EVP = explained variance percentage of the simulation over the training period.

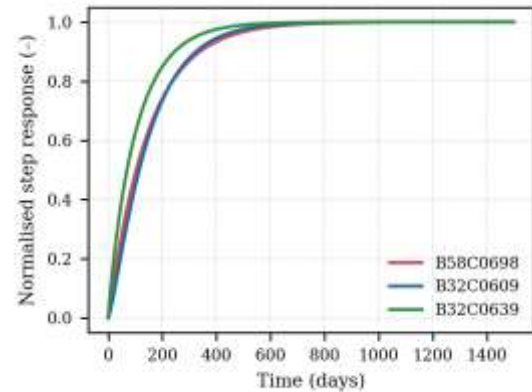


Figure 6. Normalised step responses of the calibrated TFN models

The ML permutation importance (Figure 7) gave the same picture. The 6-month cumulative surplus accounted for 91% of the importance at B58C0698, 81% at B32C0609 and 54% at B32C0639. At B32C0639 the 3-month (15%) and 12-month (13%) surpluses also contributed, which matches its shorter mean response time. Monthly P , E and the seasonal terms individually contributed less than 5%. The seasonal cycle of the heads is thus explained by the accumulated surplus rather than by the calendar.

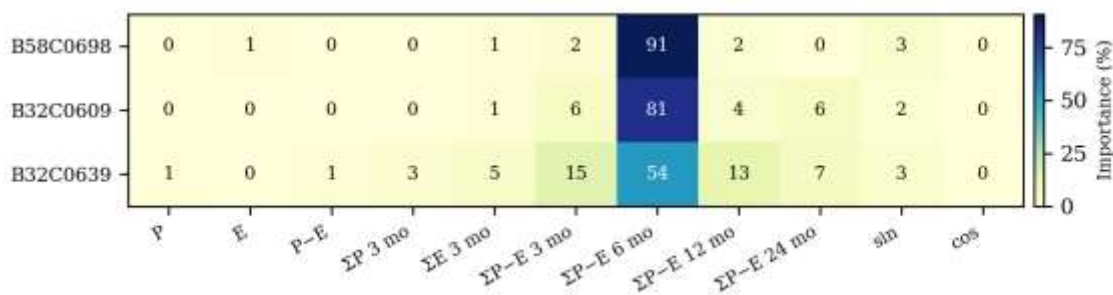


Figure 7. Normalised permutation importance of the ML inputs (mean of RF and XGBoost, test period)

4. Discussion

With forcing-only inputs, the parsimonious TFN model performed as well as the four ML models at two of the three wells, and no ML model was significantly better at any well. This agrees with the Groundwater Time Series Modelling Challenge, in which lumped-parameter models generally performed as well as artificial intelligence models (Collenteur et al., 2024). The TFN model has only six parameters, and its

structure reflects the physics of the system: the heads integrate the recharge over time. The ML models had to learn this behaviour from 200–320 monthly samples, and permutation importance showed that they did so mainly through the pre-computed 6-month cumulative surplus. In effect, the feature engineering supplied the ML models with a convolution similar to the one in the TFN model.

The LSTM, which received only the raw monthly forcing sequence, had to learn this memory itself. It was the best model at B32C0639 but among the weakest at B32C0609. This mixed result is consistent with Wunsch et al. (2021), who found that LSTM performance depends strongly on the well and on the amount of training data. Tree ensembles fitted the training data very closely, but at the two wells without data problems this closer fit did not carry over to the test periods.

The B32C0609 case shows why comparisons between models should be accompanied by an inspection of the data. The systematic negative bias of all five models after 2010 coincides with the switch to daily logger measurements. It may reflect a change in the measurement reference, a local hydrological change or a real response that the forcing does not explain. The data used here cannot distinguish between these causes. Because the TFN model is constrained by its response function, it attributed the offset to noise. The tree ensembles, which use the full non-linear input space, partly absorbed it, which explains their higher NSE at this well. Users of data-driven models should therefore not read higher skill in such a situation as evidence of better process representation.

This study has several limitations. Only three wells in a temperate, shallow-aquifer setting were analysed, and the models used only precipitation and evaporation. Abstraction, surface-water levels and land-use changes, which can influence Dutch heads, were not included. A linear recharge model was used, whereas non-linear formulations may improve TFN performance during droughts (Peterson and Western, 2014). Future work could extend the comparison to more wells with different hydrogeological settings, test hybrid models that feed TFN outputs to ML models, and use standardized groundwater indices for drought applications (Bloomfield and Marchant, 2013).

5. Conclusions

A transfer function-noise model and four machine learning models were compared for monthly groundwater level simulation at three Dutch wells, using precipitation and evaporation as the only inputs. The main conclusions are:

(1) Heads followed a regular winter–spring maximum and late-summer minimum, were best explained by the 6-month cumulative precipitation surplus, and showed no significant trend over the study periods.

(2) At two wells, all models achieved test NSE between 0.79 and 0.93, and the differences between TFN and the ML models were not significant. At the third well, all models underestimated heads by 8–12 cm after the start of daily logging. RF and XGBoost were the least affected, but their advantage over TFN was not significant at the 5% level.

(3) The TFN step responses ($t_{95} = 323$ –441 days) and the ML permutation importance consistently

identified the half-year accumulated surplus as the main control on the heads.

(4) For monthly data and forcing-only simulation, a parsimonious TFN model is a strong and interpretable benchmark. Machine learning models should be reported against such a benchmark, together with a significance test and a check of data homogeneity.

Data Availability Statement

Groundwater levels are from the Dutch DINO database (www.dinoloket.nl) and meteorological data from KNMI (www.knmi.nl), as distributed with the example data of the open-source Pastas package (<https://github.com/pastas/pastas>). Scripts are available from the corresponding author on request.

Conflict of Interest

The authors declare no conflict of interest.

References

- Bakker, M., & Schaars, F. (2019). Solving groundwater flow problems with time series analysis: you may not even need another model. *Groundwater*, 57(6), 826–833. <https://doi.org/10.1111/gwat.12927>
- Bloomfield, J.P., & Marchant, B.P. (2013). Analysis of groundwater drought building on the standardised precipitation index approach. *Hydrology and Earth System Sciences*, 17, 4769–4787. <https://doi.org/10.5194/hess-17-4769-2013>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Collenteur, R.A., Bakker, M., Caljé, R., Klop, S.A., & Schaars, F. (2019). Pastas: open source software for the analysis of groundwater time series. *Groundwater*, 57(6), 877–885. <https://doi.org/10.1111/gwat.12925>
- Collenteur, R.A., Haaf, E., Bakker, M., Liesch, T., Wunsch, A., Soonthornrangsang, J., et al. (2024). Data-driven modelling of hydraulic-head time series: results and lessons learned from the 2022 Groundwater Time Series Modelling Challenge. *Hydrology and Earth System Sciences*, 28, 5193–5208. <https://doi.org/10.5194/hess-28-5193-2024>
- Diebold, F.X., & Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.
- Gupta, H.V., Kling, H., Yilmaz, K.K., & Martinez, G.F. (2009). Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. *Journal of Hydrology*, 377, 80–91. <https://doi.org/10.1016/j.jhydrol.2009.08.003>
- Hamed, K.H., & Rao, A.R. (1998). A modified Mann-Kendall trend test for autocorrelated data. *Journal of Hydrology*, 204, 182–196.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Kendall, M.G. (1975). *Rank Correlation Methods*. Charles Griffin, London.

- Mann, H.B. (1945). Nonparametric tests against trend. *Econometrica*, 13, 245–259.
- Nash, J.E., & Sutcliffe, J.V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, 10(3), 282–290.
- Nayak, P.C., Rao, Y.R.S., & Sudheer, K.P. (2006). Groundwater level forecasting in a shallow aquifer using artificial neural network approach. *Water Resources Management*, 20, 77–90. <https://doi.org/10.1007/s11269-006-4007-z>
- Obergfell, C., Bakker, M., & Maas, K. (2019). Estimation of average diffuse aquifer recharge using time series modeling of groundwater heads. *Water Resources Research*, 55, 2194–2210. <https://doi.org/10.1029/2018WR024235>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Peterson, T.J., & Western, A.W. (2014). Nonlinear time-series modeling of unconfined groundwater head. *Water Resources Research*, 50, 8330–8355. <https://doi.org/10.1002/2013WR014800>
- Rajae, T., Ebrahimi, H., & Nourani, V. (2019). A review of the artificial intelligence methods in groundwater level modeling. *Journal of Hydrology*, 572, 336–351. <https://doi.org/10.1016/j.jhydrol.2018.12.037>
- Sahoo, S., Russo, T.A., Elliott, J., & Foster, I. (2017). Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S. *Water Resources Research*, 53, 3878–3895. <https://doi.org/10.1002/2016WR019933>
- Sen, P.K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, 63, 1379–1389.
- Taylor, R.G., Scanlon, B., Döll, P., Rodell, M., van Beek, R., Wada, Y., et al. (2013). Ground water and climate change. *Nature Climate Change*, 3, 322–329. <https://doi.org/10.1038/nclimate1744>
- Tyralis, H., Papacharalampous, G., & Langousis, A. (2019). A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water*, 11(5), 910. <https://doi.org/10.3390/w11050910>
- Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*. Springer, New York.
- von Asmuth, J.R., Bierkens, M.F.P., & Maas, K. (2002). Transfer function-noise modeling in continuous time using predefined impulse response functions. *Water Resources Research*, 38(12), 1319. <https://doi.org/10.1029/2001WR001136>
- Wunsch, A., Liesch, T., & Broda, S. (2021). Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX). *Hydrology and Earth System Sciences*, 25, 1671–1687. <https://doi.org/10.5194/hess-25-1671-2021>