

Reliability of Shallow Groundwater Levels against Low-Level Thresholds: Monte Carlo Analysis with Transfer Function-Noise Models for Three Dutch Wells

Rahleh Ahmadi¹✉, Eshagh Bandani²

¹ Contracts Affairs Expert, Sistan and Baluchestan Regional Water Company, Zahedan, Iran.

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

*Corresponding Author: Email: (r.ahmadi.sbrw@gmail.com (R. A))

Received: 11 November 2025

Accepted: 05 December 2025

Available online: 03 January 2026

Abstract: Low groundwater levels harm groundwater-dependent ecosystems, agriculture and infrastructure. Managers therefore need the probability that heads fall below critical levels, not just a simulation of past heads. This study combined transfer function-noise (TFN) models with Monte Carlo simulation to assess the reliability of monthly groundwater levels at three shallow observation wells in the Netherlands (B58C0698 near Heibloem; B32C0639 and B32C0609 near De Bilt). Pastas TFN models with a gamma response function, linear recharge and AR(1) noise explained 71–93% of the head variance. Five hundred synthetic 30-year sequences were generated for each well by block-bootstrapping hydrological years of observed daily precipitation and evaporation. Parameter uncertainty and residual noise were added stepwise, and a dry stress scenario (precipitation –5%, evaporation +10%) and uniform drawdowns of up to 0.5 m were also simulated. Failure was defined as a monthly mean head below the 10th percentile of observed monthly heads (Q10); Q05 and the mean lowest groundwater level (GLG) were used as alternative thresholds. With all uncertainty sources included, time-based reliability was 0.89–0.90 and the annual probability of at least one failure month was 0.35–0.50, consistent with the observed frequencies. Parameter uncertainty had a negligible effect, whereas residual noise reduced reliability by up to 0.03 and increased the annual failure probability by up to 0.09. Failures were concentrated in August. Resilience was low (0.33–0.43), and failure events lasted 4.4–7.1 months at most on average. The dry stress scenario reduced reliability to 0.77–0.83 and raised the annual failure probability to 0.53–0.78. A uniform drawdown of only 3–11 cm lowered reliability below 0.85, and the De Bilt well with the smallest head variability was the most sensitive. TFN-based Monte Carlo analysis is a practical tool for setting risk-based limits on groundwater abstraction.

Keywords: Groundwater level, Reliability, Transfer function-noise model, Pastas, Monte Carlo simulation, Drawdown

1. Introduction

Shallow groundwater levels control the water supply to crops and wetlands, the stability of foundations and the base flow of streams. In low-lying countries such as the Netherlands, small lowerings of the water table can damage groundwater-dependent ecosystems and agriculture. Dry summers such as 2018 have drawn attention to groundwater drought (Van Loon, 2015; Philip et al., 2020). Water managers therefore need to know how often heads fall below critical levels and how this frequency changes under abstraction and climate change. In the Netherlands, the mean lowest and highest groundwater levels (GLG and GHG) are standard descriptors of the regime (Knotters and van Walsum, 1997; von Asmuth and Knotters, 2004).

Transfer function-noise (TFN) models relate observed heads to precipitation and evaporation through an impulse response function and represent the unexplained part by a noise model (von Asmuth et al., 2002). They are parsimonious, can be calibrated on irregular observations and are available in the open-source Pastas package (Collenteur et al., 2019). Beyond simulation, TFN models have been used to estimate recharge (Oberghell et al., 2019), to replace groundwater flow models for some tasks (Bakker and Schaars, 2019),

and in the 2022 groundwater modelling challenge (Collenteur et al., 2024). Because they run very quickly, they are well suited to Monte Carlo analysis.

The reliability, resilience and vulnerability (RRV) framework of Hashimoto et al. (1982) quantifies how often a system fails, how quickly it recovers and how severe failures are. It has been widely applied to water supply systems (Kjeldsen and Rosbjerg, 2004), but rarely to groundwater levels. Groundwater drought indices based on standardised heads describe the state of the system (Bloomfield and Marchant, 2013; Marchant and Bloomfield, 2018), but they do not give the probability of crossing a management threshold under uncertain future weather. In addition, the relative importance of climate variability, model parameter uncertainty and model structural error (residual noise) for such probabilities is rarely quantified.

The objective of this study was to develop and apply a TFN-based Monte Carlo framework for the reliability of monthly groundwater levels. The specific aims were to (i) calibrate and validate TFN models for three Dutch wells; (ii) estimate reliability, resilience, vulnerability and annual failure probabilities for low-level thresholds; (iii) separate the contributions of climate variability, parameter uncertainty and residual noise; and (iv)

evaluate a dry climate stress test and uniform drawdowns such as those caused by abstraction.

2. Materials and Methods

2.1. Wells and Data

Three observation wells from the Dutch national subsurface database (DINO, TNO), distributed as example data with the Pastas package, were used (Figure 1; Table 1). Well B58C0698 near Heibloem (Limburg) was combined with daily precipitation from Heibloem and Makkink reference evaporation from Maastricht. Wells B32C0639 and B32C0609 near De Bilt (Utrecht) were combined with daily precipitation and Makkink evaporation from KNMI station De Bilt (260). Most heads were measured about twice a month. For B32C0609, the record was truncated at December 2009, before daily logger measurements began, because a previous analysis showed a bias after that change. Monthly mean heads were computed from all observations in each month.

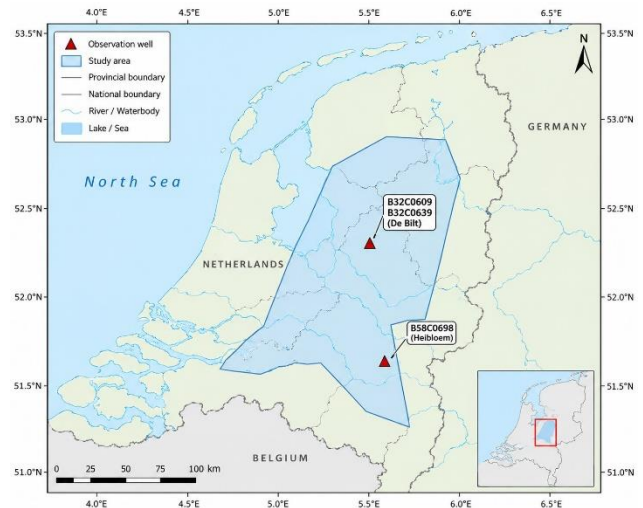


Figure 1. Approximate locations of the observation wells

Table 1. Characteristics of the wells, groundwater regime indicators and thresholds (heads in m relative to NAP)

Well	Location	Period	Obs.	Months	Mean	SD	Min / max	GHG	GLG	Q10	Q05	Trend annual min., cm/decade (p)
B58C0698	Heibloem (Limburg)	1985-11 – 2015-06	644	341	27.90	0.42	26.71 / 28.89	28.44	27.31	27.34	27.21	+6.4 (0.36)
B32C0639	De Bilt (Utrecht)	1982-01 – 2005-10	532	280	1.36	0.20	0.95 / 1.96	1.64	1.14	1.13	1.10	-0.8 (0.83)
B32C0609	De Bilt (Utrecht)	1982-01 – 2009-12	607	316	1.18	0.32	0.51 / 2.12	1.52	0.91	0.78	0.70	+7.5 (0.28)

Statistics of monthly mean heads; GHG/GLG = mean of the three highest/lowest observations per hydrological year (April–March); Q10/Q05 = 10th/5th percentiles of observed monthly heads.

2.2. Transfer Function-Noise Model

The head h was modelled as the convolution of the recharge R with an impulse response function θ , plus a base level d and a noise term n (von Asmuth et al., 2002; Collenteur et al., 2019):

$$h(t) = d + \int_0^t R(\tau) \theta(t - \tau) d\tau + n(t), \quad R = P + fE \quad (1)$$

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

where P and E are daily precipitation and evaporation, f is an evaporation factor, and A , n and a are the gain, shape and scale of the gamma response. The residuals were modelled by an AR(1) process with characteristic time α . The six parameters were estimated with the Pastas least-squares solver (version 2.0) using the whole record and a 10-year warm-up. For the Monte Carlo runs, the model was re-implemented as a direct convolution of daily recharge with the Pastas block response; it reproduced the Pastas simulation to within 1.6 mm.

2.3. Monte Carlo Simulation

Four simulation sets (S0–S3) of 500 sequences each were generated per well. Each sequence consisted of 40 hydrological years (April–March), of which the first 10 were discarded as warm-up, leaving 30 years (360

months) for evaluation. Synthetic climate was created by drawing complete hydrological years of observed daily precipitation and evaporation with replacement. This block bootstrap preserves the within-year sequence and the joint behaviour of the two variables (Efron and Tibshirani, 1993; Lall and Sharma, 1996).

S0 (climate only) used the optimal parameters and no noise. **S1** added parameter uncertainty: each sequence used a parameter set drawn from the truncated multivariate normal distribution defined by the estimated covariance matrix. **S2** added residual noise as a daily AR(1) process with the fitted characteristic time and the standard deviation of the model residuals σ_r :

$$\varepsilon_t = \phi \varepsilon_{t-1} + \sigma_r \sqrt{1 - \phi^2} z_t, \quad \phi = e^{-1/\alpha} \quad (3)$$

where z_t is a standard normal deviate. **S3** was a dry stress test: S2 with precipitation reduced by 5% and evaporation increased by 10% on every day. S3 is a sensitivity case in the spirit of scenario-neutral analysis (Prudhomme et al., 2010; Brown et al., 2012), not a climate projection. Finally, the S2 heads were lowered uniformly by 0–0.5 m in 0.05 m steps to represent the drawdown caused by abstraction or drainage.

2.4. Thresholds and Performance Indices

A month was a failure if the simulated monthly mean head was below a threshold h_c . The main threshold was

the 10th percentile of observed monthly heads (Q10), and the 5th percentile (Q05) and GLG were used for comparison. For each sequence, time-based reliability R_t , resilience Res and vulnerability Vul were computed (Hashimoto et al., 1982):

$$R_t = 1 - \frac{N_f}{N}, \quad Res = \frac{N_{fs}}{N_f} \quad (4)$$

$$Vul = \frac{1}{M} \sum_{j=1}^M \max_{t \in j} (h_c - h_t) \quad (5)$$

where N_f is the number of failure months among $N = 360$, N_{fs} the number of transitions from failure to non-failure, and M the number of failure events; vulnerability is the mean of the maximum deficit (cm) of each event. In addition, the maximum event duration D_{max} , the annual failure probability P_a (the fraction of years with at least one failure month) and the probability of no

failure in the first 10 years were computed. The indices were averaged over the 500 sequences. The same statistics were computed for the observed and the historical simulated heads as a check. All analyses were coded in Python with Pastas, NumPy, SciPy and pandas.

3. Results and Discussion

3.1. Groundwater Regime and Model Performance

Mean monthly heads were 27.90 m at B58C0698, 1.36 m at B32C0639 and 1.18 m at B32C0609 (Table 1). The seasonal amplitude, expressed as GHG – GLG, was 1.13, 0.49 and 0.61 m, respectively. The lowest heads occurred in August–September. The Q10 thresholds lay close to the GLG at B58C0698 and B32C0639 but 0.13 m below it at B32C0609. No significant trend was found in annual minimum heads ($p \geq 0.28$).

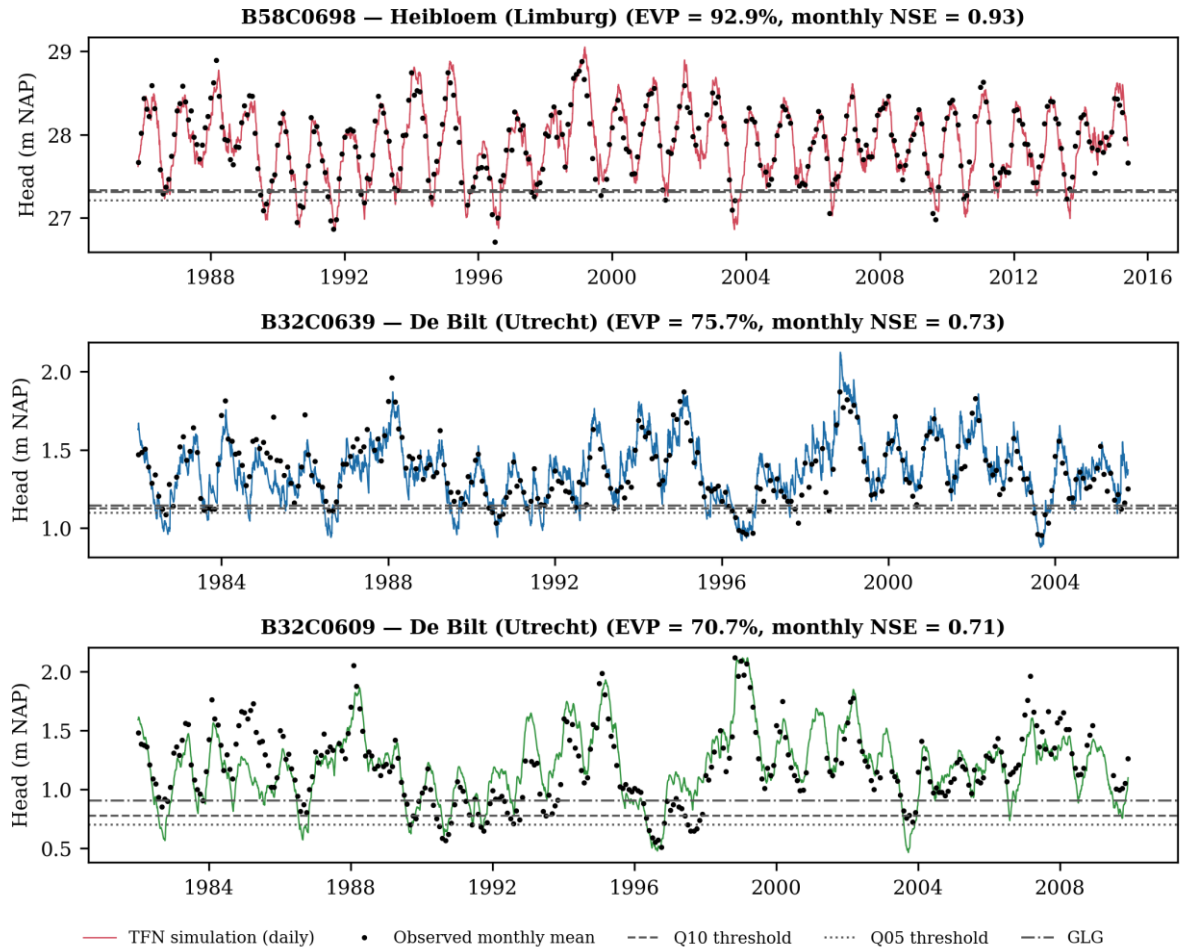


Figure 2. Observed monthly mean heads, TFN simulations and the low-level thresholds

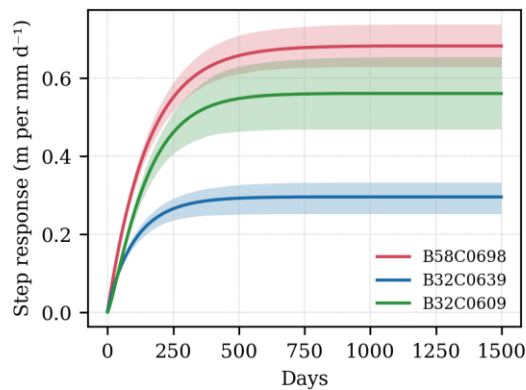


Figure 3. Step responses of the TFN models with 90% bands from parameter uncertainty

Table 2. Calibrated TFN parameters ($\pm$ standard error) and goodness of fit

Well	A (m per mm d ⁻¹)	n (-)	a (d)	f (-)	d (m)	α (d)	t_{95} (d)	EVP (%)	NSE monthly	σ_r (cm)
B58C0698	0.682 $\pm$ 0.036	1.02 $\pm$ 0.02	150 $\pm$ 11	-1.27 $\pm$ 0.06	27.88 $\pm$ 0.07	50 $\pm$ 6	456	92.9	0.93	11.5
B32C0639	0.296 $\pm$ 0.025	0.91 $\pm$ 0.03	119 $\pm$ 15	-0.58 $\pm$ 0.07	0.96 $\pm$ 0.05	54 $\pm$ 7	335	75.7	0.73	10.2
B32C0609	0.560 $\pm$ 0.058	1.31 $\pm$ 0.03	115 $\pm$ 12	-0.60 $\pm$ 0.06	0.43 $\pm$ 0.12	196 $\pm$ 42	411	70.7	0.71	17.4

The TFN models explained 70.7–92.9% of the head variance, with monthly NSE of 0.71–0.93 (Table 2; Figure 2). The fit was best at B58C0698, where heads follow a regular seasonal cycle. The gains were well constrained (standard errors 5–11% of the estimates), and the step responses reached 95% of their final values after 335–456 days (Figure 3). The evaporation factors of -1.27 to -0.58 indicate that evaporation reduces

recharge less than one-to-one at the De Bilt wells and more than one-to-one at Heibloem. The residual standard deviation was 10–17 cm, and the noise characteristic time was 50–196 days. The residuals are thus persistent, and noise will matter for monthly means.

3.2. Reliability under Climate Variability and Model Uncertainty

Table 3. Reliability indices of monthly heads for 500 synthetic 30-year sequences per scenario (Q10 threshold unless stated)

Well	Scenario	R _t	R _t 5th pct.	P _a	Res	Vul (cm)	D _{max} (mo)	Events	R _t Q05	P _a Q05	R _t GLG	P _a GLG
B58C0698	Climate	0.910	0.869	0.459	0.44	19.4	4.2	13.8	0.953	0.293	0.918	0.428
	+ Parameters	0.910	0.867	0.459	0.44	19.4	4.1	13.8	0.952	0.295	0.918	0.429
	+ Noise	0.904	0.861	0.480	0.43	21.0	4.4	14.7	0.947	0.319	0.912	0.452
	Dry stress	0.769	0.706	0.784	0.30	32.5	6.1	24.5	0.845	0.648	0.782	0.766
	Observed	0.900	–	0.556	–	–	–	–	0.950	–	–	–
B32C0639	Climate	0.918	0.875	0.425	0.45	8.7	5.0	13.0	0.940	0.340	0.901	0.475
	+ Parameters	0.918	0.867	0.425	0.46	8.6	5.0	13.0	0.939	0.341	0.900	0.475
	+ Noise	0.888	0.836	0.504	0.43	10.6	5.9	16.9	0.914	0.419	0.870	0.554
	Dry stress	0.821	0.750	0.678	0.37	12.6	7.3	23.3	0.857	0.596	0.796	0.725
	Observed	0.900	–	0.417	–	–	–	–	0.950	–	–	–
B32C0609	Climate	0.933	0.889	0.265	0.35	12.4	5.3	8.1	0.964	0.167	0.851	0.497
	+ Parameters	0.932	0.875	0.267	0.36	12.2	5.6	8.1	0.963	0.165	0.852	0.491
	+ Noise	0.902	0.828	0.352	0.33	16.6	7.1	11.0	0.937	0.250	0.818	0.540
	Dry stress	0.827	0.728	0.528	0.28	20.0	9.7	16.6	0.881	0.410	0.710	0.719
	Observed	0.899	–	0.280	–	–	–	–	0.956	–	–	–

R_t = time-based reliability; P_a = annual failure probability; Res = resilience; Vul = mean maximum deficit per event; D_{max} = longest event per sequence; Events = number of failure events per 30 years.

With climate variability only (S0), the time-based reliability for the Q10 threshold was 0.910–0.933 and the annual failure probability 0.27–0.46 (Table 3; Figure 5). Adding parameter uncertainty (S1) changed these values by less than 0.003, because the parameters were well constrained by 24–30 years of data. Adding residual noise (S2) had a larger effect: reliability decreased to 0.888–0.904, the annual failure probability increased to

0.35–0.50, and vulnerability increased by 1.6–4.3 cm. The effect of noise was largest at B32C0609, which had the largest residual standard deviation and the longest noise memory. Model structural error, rather than parameter uncertainty, is therefore the main model-related source of uncertainty for threshold exceedance.

The S2 ensembles reproduced the historical behaviour well. By construction, the observed reliability

for Q10 was 0.90, and the S2 values were 0.888–0.904. The observed annual failure frequencies (0.28–0.56) were within 0.1 of the S2 values (0.35–0.50). The seasonal distribution of the simulated heads matched the

observed monthly medians and percentiles (Figure 4). The annual minimum heads of the observations lay close to the S2 frequency curves (Figure 6).

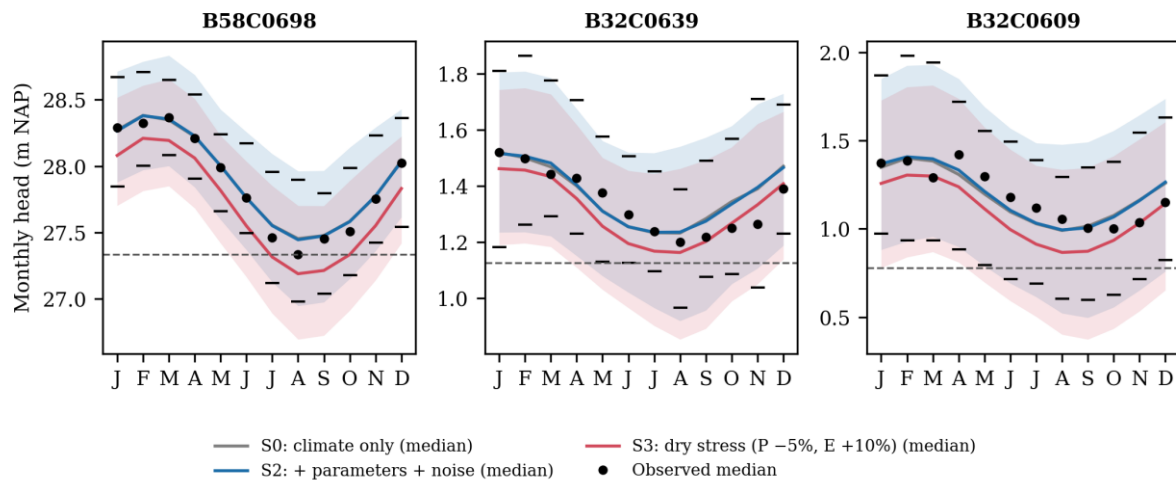


Figure 4. Seasonal distribution of monthly heads: ensemble medians and 5–95% ranges compared with observed medians (dots) and 5th/95th percentiles (bars); dashed line = Q10

Failures were persistent. Resilience was only 0.33–0.43, so a failure month was followed by another failure month in more than half of the cases. The longest event in a 30-year sequence lasted 4.4–7.1 months on average. Vulnerability was largest at B58C0698 (21 cm), which has the largest seasonal amplitude. Almost no sequence was failure-free over its first 10 years (probability $\leq$

0.04), which is expected for a threshold exceeded in 10% of months. The stricter Q05 threshold reduced the annual failure probability to 0.25–0.42. Using the GLG as threshold gave results similar to Q10 at B58C0698 and B32C0639 and much lower reliability at B32C0609 (0.818), where the GLG lies higher in the distribution of heads.

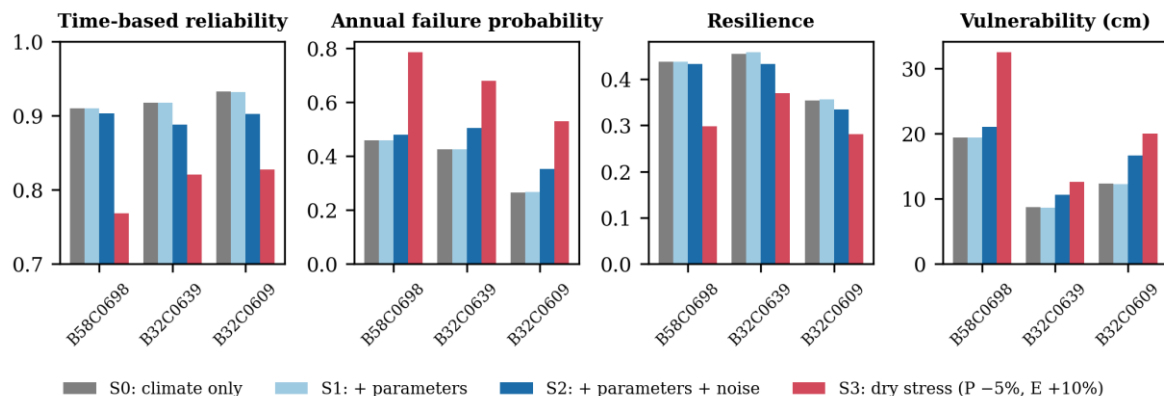


Figure 5. Reliability, annual failure probability, resilience and vulnerability for the four simulation sets (Q10 threshold)

3.3. Seasonality of Failure

The monthly failure probability peaked in August at all wells (Figure 7). At B58C0698, failures were almost entirely confined to June–November, with 0.38 of Augusts below Q10. At the De Bilt wells, failures were spread over more months, and B32C0609 had a relatively high probability of low heads in autumn and

early winter. This reflects its longer memory, which delays the recovery after dry summers. The observed monthly failure frequencies followed the same pattern. The simulated peak at B58C0698 was somewhat lower than observed (0.38 against 0.48 in August), probably because each observed value is based on only about 30 Augusts.

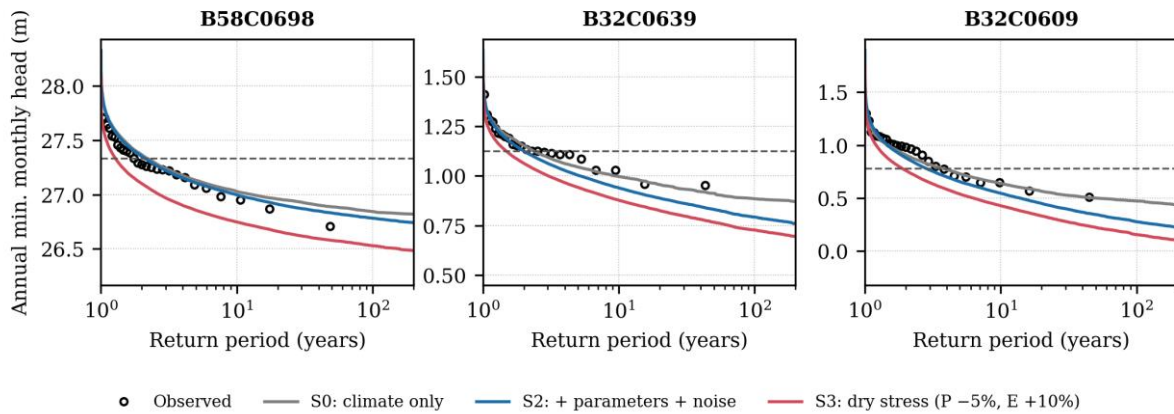


Figure 6. Return-period plots of annual minimum monthly heads: observations and ensembles S0, S2 and S3; dashed line = Q10

3.4. Dry Stress Test and Drawdown

The dry stress scenario (S3) lowered the mean simulated head by 6–21 cm. It reduced reliability to 0.769–0.827 and raised the annual failure probability to 0.53–0.78 (Table 3). At B58C0698, where the gain is largest, the annual failure probability rose from 0.48 to 0.78, and vulnerability increased from 21 to 32 cm. The annual minimum head with a 10-year return period under S3 had a return period of 24–24 years at the De Bilt wells and 174 years at B58C0698 under S2 (Figure 6). A moderate shift in the water balance therefore makes rare low levels common. This supports the concern about increasing summer drought in the Netherlands (Philip et al., 2020).

Uniform drawdown reduced reliability almost linearly at first (Figure 7). A reliability of 0.85 was reached with a drawdown of 11 cm at B58C0698, 8 cm at B32C0609 and only 3 cm at B32C0639. The difference follows from the head variability: the smaller the standard deviation of heads near the threshold, the more months cross it for a given lowering. A drawdown of 10 cm increased the annual failure probability to 0.62 (B58C0698), 0.77 (B32C0639) and 0.50 (B32C0609). Such curves can be used to set site-specific maximum drawdowns for abstraction permits from an explicit reliability target, instead of a uniform allowable lowering.

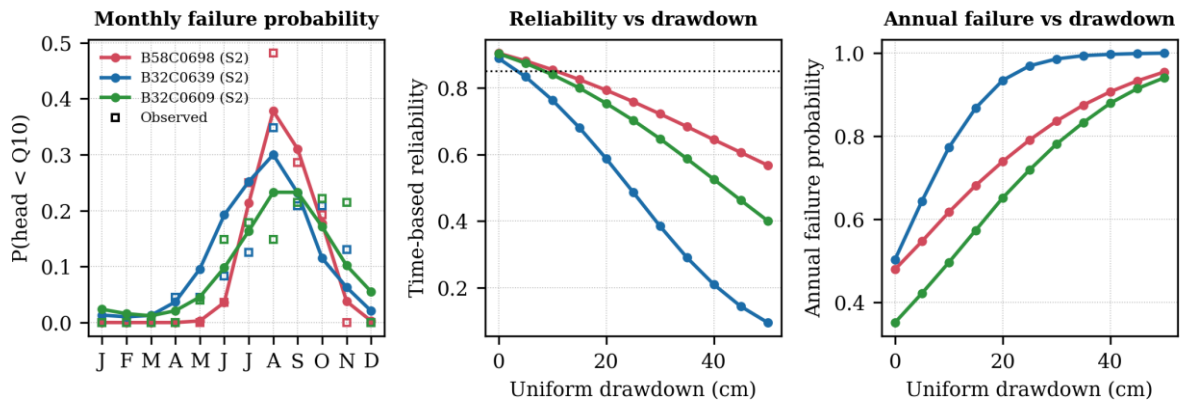


Figure 7. Monthly probability of heads below Q10 (S2, lines; observed, squares) and reliability and annual failure probability versus uniform drawdown

3.5. Limitations

The TFN models use a linear recharge concept and a single response function; non-linear recharge models could represent dry years better (Peterson and Western, 2014). The block bootstrap cannot generate years drier than the observed ones or changes in the sequence of wet and dry years, and the dry stress test applied a uniform change to all days. The thresholds were statistical, not ecological or agronomic; site-specific critical levels would be needed in practice. Drawdown was represented as a uniform shift, whereas real abstraction produces a time-varying drawdown that could be

modelled with an additional well stress in Pastas. Finally, only three wells in the Netherlands were analysed, and the observed thresholds depend on the record period.

4. Conclusions

A Monte Carlo framework based on Pastas transfer function-noise models was used to assess the reliability of monthly groundwater levels at three Dutch wells. The main conclusions are:

(1) The TFN models explained 71–93% of the head variance, and their ensembles reproduced the observed seasonal distribution and frequency of low heads.

(2) With climate variability, parameter uncertainty and residual noise, the time-based reliability for the Q10 threshold was 0.89–0.90 and the annual failure probability 0.35–0.50. Residual noise, not parameter uncertainty, was the main model-related source of uncertainty.

(3) Low-head events were persistent (resilience 0.33–0.43) and concentrated in August–September.

(4) A dry stress of –5% precipitation and +10% evaporation raised the annual failure probability to 0.53–0.78, and a uniform drawdown of 3–11 cm reduced reliability below 0.85.

(5) Reliability–drawdown curves derived from TFN ensembles provide a transparent basis for risk-based abstraction limits.

Data Availability Statement

The groundwater levels (DINO, TNO) and meteorological data (KNMI) are 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>
- Brown, C., Ghile, Y., Lavery, M., & Li, K. (2012). Decision scaling: Linking bottom-up vulnerability analysis with climate projections in the water sector. *Water Resources Research*, 48, W09537. <https://doi.org/10.1029/2011WR011212>
- 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>
- Efron, B., & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.
- Hashimoto, T., Stedinger, J.R., & Loucks, D.P. (1982). Reliability, resiliency, and vulnerability criteria for water resource system performance evaluation. *Water Resources Research*, 18(1), 14–20. <https://doi.org/10.1029/WR018i001p00014>
- Kjeldsen, T.R., & Rosbjerg, D. (2004). Choice of reliability, resilience and vulnerability estimators for risk assessments of water resources systems. *Hydrological Sciences Journal*, 49(5), 755–767. <https://doi.org/10.1623/hysj.49.5.755.55136>
- Knotters, M., & van Walsum, P.E.V. (1997). Estimating fluctuation quantities from time series of water-table depths using models with a stochastic component. *Journal of Hydrology*, 197, 25–46.
- Lall, U., & Sharma, A. (1996). A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research*, 32(3), 679–693. <https://doi.org/10.1029/95WR02966>
- Marchant, B.P., & Bloomfield, J.P. (2018). Spatio-temporal modelling of the status of groundwater droughts. *Journal of Hydrology*, 564, 397–413.
- Oberghell, 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>
- 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>
- Philip, S.Y., Kew, S.F., van der Wiel, K., Wanders, N., & van Oldenborgh, G.J. (2020). Regional differentiation in climate change induced drought trends in the Netherlands. *Environmental Research Letters*, 15, 094081. <https://doi.org/10.1088/1748-9326/ab97ca>
- Prudhomme, C., Wilby, R.L., Crooks, S., Kay, A.L., & Reynard, N.S. (2010). Scenario-neutral approach to climate change impact studies: Application to flood risk. *Journal of Hydrology*, 390, 198–209. <https://doi.org/10.1016/j.jhydrol.2010.06.043>
- Van Loon, A.F. (2015). Hydrological drought explained. *WIREs Water*, 2, 359–392. <https://doi.org/10.1002/wat2.1085>
- von Asmuth, J.R., & Knotters, M. (2004). Characterising groundwater dynamics based on a system identification approach. *Journal of Hydrology*, 296, 118–134.
- 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>