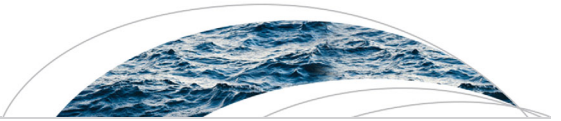
AGU PUBLICATIONS

Water Resources Research



RESEARCH ARTICLE

10.1002/2016WR019704

Key Points:
- Length calibration period more important than number of observations for calibrating time series models of groundwater dynamics
- Credible intervals can be reduced to 10% and prediction intervals to 50% of head range with 10–20 years of observations
- Required length calibration period stronger related to decay time of noise than to response time of system

Supporting Information:
- Supporting Information S1

Correspondence to:
M. Bakker,
Mark.Bakker@tudelft.nl

Citation:
van der Spek, J. E., and M. Bakker (2017), The influence of the length of the calibration period and observation frequency on predictive uncertainty in time series modeling of groundwater dynamics, *Water Resour. Res.*, *53*, 2294–2311, doi:10.1002/2016WR019704.

Received 26 AUG 2016
Accepted 23 FEB 2017
Accepted article online 27 FEB 2017
Published online 22 MAR 2017

The influence of the length of the calibration period and observation frequency on predictive uncertainty in time series modeling of groundwater dynamics

Joanne E. van der Spek[1] and Mark Bakker[1]

[1]Water Resources Section, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, the Netherlands

Abstract The influence of the length of the calibration period and observation frequency on the predictive uncertainty in time series modeling of groundwater dynamics is investigated. Studied series are from deltaic regions with predominantly shallow groundwater tables in a temperate maritime climate where heads vary due to precipitation and evaporation. Response times vary over a wide range from ∼60 to ∼1200 days. A Transfer Function-Noise model is calibrated with the Markov Chain Monte Carlo method to both synthetic series and measured series of heads. The model fit and uncertainty are evaluated for various calibration periods and observation frequencies. It is often assumed that the required length of the calibration period is related to the response time of the system. In this study, no strong relationship was observed. Results indicate, however, that the required length of the calibration period is related to the decay time of the noise. Furthermore, the length of the calibration period was much more important than the total number of observations. For the measured series, the credible intervals could commonly be reduced to ∼10% of the measured head range and the prediction intervals to ∼50% of the measured head range with calibration periods of 20 years with approximately two observations per month.

1. Introduction

Over the past decades, time series analysis has become an important tool in groundwater hydrology. Time series models are frequently used to characterize groundwater dynamics [e.g., *Finke et al*., 2004; *Von Asmuth and Knotters*, 2004], decompose head fluctuations into the impacts of multiple stresses [e.g., *Shapoori et al*., 2015a, 2015b; *Von Asmuth et al*., 2008], and estimate unobserved heads [e.g., *Knotters and Van Walsum*, 1997]. Time series analysis is highly data-driven, relying on the correspondence between input and output series. One of the major questions that arises when developing a time series model is how long and how frequent the head needs to be measured before reliable results can be expected.

Time series analysis has been widely applied since its extensive description by *Box and Jenkins* [1970]. Transfer Function-Noise (TFN) models are a type of time series models often used in groundwater hydrology [e.g., *Knotters and Van Walsum*, 1997; *Von Asmuth et al*., 2002; *Berendrecht et al*., 2003; *Finke et al*., 2004; *Obergfell et al*., 2013; *Peterson and Western*, 2014] and consist of a deterministic part and a stochastic part. Output series are modeled by transformation of input series, while the autocorrelation in the differences between the observed and simulated output series is described by a noise model. TFN models owe their popularity to their relatively easy construction and accurate results [*Von Asmuth et al*., 2008]. *Von Asmuth et al*. [2002] presented a TFN model for groundwater dynamics based on predefined response functions. The head is simulated as the sum of the responses to past stresses like precipitation and evaporation, and a constant base elevation. The responses to stresses are calculated by convolution of observed stresses with appropriate response functions. *Von Asmuth et al*. [2002] adopted a scaled-gamma function as response function for precipitation and evaporation. This function has three parameters giving it a flexible shape varying from exponential to skewed bell-shaped, the latter representing a delayed response. Noise decay is approximated as exponential, which makes the model suitable for use with unevenly spaced data [*Von Asmuth and Bierkens*, 2005; *Yang et al*., 2007].

Assessment of parameter and predictive uncertainty are important aspects of hydrological modeling [*Schoups and Vrugt*, 2010; *Thyer et al*., 2009]. Uncertainty in model predictions is caused by input uncertainty,

for example caused by measurement errors in precipitation and evaporation data, output uncertainty, for example caused by errors in observed heads, structural uncertainty caused by simplifications in the model structure, and parameter uncertainty [*Renard et al*., 2010]. Input uncertainty and output uncertainty can collectively be called data uncertainty [*Liu and Gupta*, 2007]. The values of the parameters of a TFN model are estimated by calibrating the model to a set of observations. Simplifications in the model structure and errors in the calibration data make it impossible to determine the parameter values exactly. Bayesian statistics offers the possibility to incorporate parameter uncertainty in the calibration process and to estimate the posterior probability distributions of the parameters based on prior knowledge and the information contained in the data [e.g., *Box and Tiao*, 1992]. The posterior distributions can be approximated with Markov Chain Monte Carlo (MCMC) methods [e.g., *Vrugt*, 2016]. These methods are widely used in hydrological studies [e.g., *Kuczera and Parent*, 1998; *Thyer and Kuczera*, 2003; *Marshall et al*., 2004; *Yang et al*., 2007; *Vrugt et al*., 2008; *Schoups and Vrugt*, 2010], including studies concerning groundwater modeling [*Fu and Gómez-Hernández*, 2009; *Hassan et al*., 2009; *Lu et al*., 2012; *Laloy et al*., 2013].

Uncertainty analysis of hydrological models has been an important research topic for many decades (e.g., see reviews by *Beck* [1987], *Liu and Gupta* [2007], *McMillan et al*. [2011], and *Moradkhani and Sorooshian* [2009]). One of the important issues is the calibration data requirement. *Sorooshian et al*. [1983] suggest that at least 1 year of calibration data is needed to effectively calibrate a conceptual rainfall-runoff model, but they state that the information contained in the calibration data is more important than the length of the data set. *Yapo et al*. [1996] found that with about 8 years of calibration data, the calibration results of a conceptual rainfall-runoff flood forecasting model are relatively insensitive to the selected period. *Perrin et al*. [2007] show that robust estimates of parameter values of a rainfall-runoff model can be obtained with 350 observations sampled from a longer period. Furthermore, the specific temporal hydrological circumstances during the calibration period are important [*Juston et al*., 2009; *Yapo et al*., 1996; *Knotters*, 2001]. In groundwater hydrology it is often assumed that the length of the calibration period required to calibrate a time series model of groundwater dynamics is related to the response time, which is the period over which a change in stress influences the head [e.g., *Knotters and Van Walsum*, 1997]. *Knotters and van Walsum* [1997] reported the results of two observation wells with shallow water table depths in the Netherlands and concluded that they only needed 4 years of calibration data to satisfactorily describe the groundwater dynamics caused by excess precipitation. Furthermore, it is sometimes thought that a high observation frequency, easily attainable with modern pressure transducers, can compensate for a short calibration period.

Understanding of the impact of the available calibration data on the predictive uncertainty in time series modeling of groundwater dynamics is currently lacking. The objective of this study is to evaluate how the length of the calibration period and the observation frequency influence the predictive uncertainty in time series modeling of groundwater dynamics and to identify the main factors that affect the uncertainty. This paper is restricted to groundwater dynamics caused by precipitation and evaporation. Both synthetic series and long observation series are evaluated. A linear TFN model, described in section 2, is calibrated while the calibration periods and observation frequencies are varied. The calibration method is explained in section 3. The methods used to evaluate the uncertainty of the results are described in section 4, followed by a description of the study area in section 5. Results for synthetic series of heads, generated with a wide range of parameter values, are presented in section 6. Results for 18 observed series of heads with a wide range of response times are presented in section 7. A discussion and conclusions are presented in sections 8 and 9, respectively.

2. Transfer Function-Noise Model

The TFN model used in this research is the same as models used by, e.g., *Von Asmuth et al*. [2002], *Obergfell et al*. [2013], and *Peterson and Western* [2014]. Observed heads $h_o(t)$ [L] are modeled as the sum of simulated heads $h(t)$ [L] and a series of residuals $r(t)$ [L]

$$h_o(t) = h(t) + r(t). \tag{1}$$

The simulated heads are obtained through convolution of the groundwater recharge and a response function, plus a constant base level

$$h(t)=d+\int_{-\infty}^{t} N(\tau)\theta(t-\tau)d\tau, \tag{2}$$

where $N(t)$ [L/T] is the groundwater recharge, $\theta(t)$ is the response function, and d [L] is the base level. The recharge is calculated as

$$N(t)=P(t)-fE(t), \tag{3}$$

where P [L/T] is precipitation, E [L/T] is reference evaporation, and f is a factor that scales the reference evaporation to the actual evaporation and is called the evaporation factor here.

Following *Von Asmuth et al.* [2002], a scaled gamma function with shape parameters A [T], a [T], and n is used for the response function

$$\theta(t)=A\frac{t^{n-1}e^{-t/a}}{\Gamma(n)a^n} \quad (0 \le t \le t_p), \tag{4}$$

$$\theta(t)=0 \quad (t < 0 \text{ and } t > t_p). \tag{5}$$

The scaled gamma function (equation (4)) approaches 0 for t going to infinity, but is truncated at time t_p, which is the time at which $100p$ percent of the response to recharge has taken place. Here p is set to 0.999. A startup period with at least length t_p must be used to include the impact of the stress on the heads at the beginning of the simulation period.

The residuals $r(t)$ are calculated as the differences between the observed and simulated heads

$$r(t)=h_o(t)-h(t). \tag{6}$$

A noise model is used to calculate the random component of the residuals, called the innovations. The noise model used here is equal to the noise models used by, e.g., *Von Asmuth et al.* [2002], *Yang et al.* [2007], *Peterson and Western* [2014], and *Shapoori et al.* [2015a]. Noise decay is approximated as exponential, such that the noise model is suitable for use with unevenly spaced head observations. More complex noise models are not easily adapted to unevenly spaced data. The innovations $v(t)$ [L] are computed as

$$v(t)=r(t)-e^{-\Delta t/\alpha}r(t-\Delta t), \tag{7}$$

where α [T] determines the decay rate of the noise. The exponential noise model largely removes autocorrelation in the residuals of many times series [e.g., *Von Asmuth and Bierkens*, 2005; *Von Asmuth et al.*, 2002], including the ones considered in this paper.

The standard deviation of the innovations σ_v [L] depends on the time step between them [*Von Asmuth and Bierkens*, 2005; *Yang et al.*, 2007]. The larger the time step, the larger the standard deviation. If the time step becomes very large, the standard deviation of the innovations approaches the standard deviation of the residuals, which is assumed to be constant. The relationship between the standard deviation of the innovations and the time step Δt is (Appendix A)

$$\sigma_v(\Delta t)=\sigma_r\sqrt{1-e^{-2\Delta t/\alpha}}, \tag{8}$$

where σ_r [L] is the standard deviation of the residuals.

3. Calibration

The model is calibrated to observed heads h_o. Measurements of precipitation and evaporation are available on a daily basis, while head observations are commonly available for unevenly spaced moments in time t_i ($i=1, 2, \ldots, M$, where M is the number of observations). The head h is simulated on a daily basis and compared to the available observations in order to estimate the probability distributions of in total seven parameters: f, d, A, a, n, α, and σ_r. It is noted that in earlier studies by, e.g., *Von Asmuth et al.* [2002] and *Peterson and Western* [2014] the model is calibrated for the parameters f, A, a, n, and α, while the values of d and σ_r are estimated after the calibration. The advantage of including d as one of the calibration parameters is

Table 1. Prior Distributions of Parameters, Used for Synthetic and Measured Data

Parameter	Prior
f	$U[0, 2]$
d_* (m)	$U[-20, 150]$
A (day)	$U[0, 10000]$
a (day)	$U[1e-9, 2000]$
n	$U[0, 5]$
α (day)	$U[1e-9, 1000]$
σ_* (m)	$U[1e-9, 1]$

that the uncertainty of this parameter is estimated directly. Estimation of the distribution of σ_r is common when applying Markov Chain Monte Carlo (MCMC) to the likelihood function.

Bayes' theorem states that the posterior probability of the parameters given the data is proportional to the product of the likelihood L of the parameters given the data and the prior probability of the parameters [*Box and Tiao*, 1992]

$$p(\phi|D) \propto L(\phi|D)p(\phi), \tag{9}$$

where ϕ stands for the parameters and D for the data. A normal likelihood function is used with uniform prior distributions for all parameters. The likelihood function for the innovations $v(t)$ is written as

$$L=\prod_{i=2}^{M}\frac{1}{\sigma_v(\Delta t_{i-1})\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{v(t_i)}{\sigma_v(\Delta t_{i-1})}\right)^2}, \tag{10}$$

where $v(t_i)$ is calculated with equations (2), (6), and (7), and $\sigma_v(\Delta t_{i-1})$ is the standard deviation of $v(t_i)$ (equation (8)), where Δt_{i-1} is the time step between t_{i-1} and t_i. In the remaining part of this paper the log-likelihood function is used.

Parameter sets are sampled from the posterior probability distributions using PyMC2 [*Patil et al*., 2010], a Python package for MCMC. Markov chains are obtained for all parameters. The distribution of the values in these chains converges to a stationary distribution that equals the posterior probability distribution (for an explanation of MCMC see, e.g., *Dunn and Shultis*, [2012]). The shapes of the posterior distributions are evaluated to assess whether convergence to a stationary distribution was achieved and the distributions are checked for multimodality. For the synthetic time series, the parameter values that give the largest log likelihood are compared to the true parameter values. Additionally, reproducibility of the MCMC results is ascertained by repetition of the parameter sampling for a number of time series.

Correlation between parameters may impede good mixing of MCMC chains. Two modifications, explained in Appendix B, are used to reduce problems with mixing of the MCMC chains due to correlation between the evaporation factor f and base elevation d, and between the decay factor of the noise α and the standard deviation of the residuals σ_r. The distributions of the altered parameters d^* [L] and σ^* [L] are sampled and afterward transformed to those of d and σ_r. In all cases the same prior distributions (Table 1) are used.

In total 105,000 samples are drawn. The chains are started at the Maximum A Posteriori (MAP) estimates (calculated with PyMC2's MAP function) of the parameter values to reduce burn-in. Nonetheless, the first 5000 values are discarded, because the first parts of the chains seem to be less well mixed. A thinning factor of 10 is used, resulting in chains of 10,000 values.

A start-up period of 1.5 times the $t_{0.999}$ of the MAP estimate of the response function is included before the beginning of the observed series of heads, such that sufficient startup data are available for sampled parameter sets with varying response times. The entire series of precipitation and evaporation are used to estimate the MAP values.

4. Evaluation of Uncertainty

Uncertainty in simulated heads is caused by parameter uncertainty, data uncertainty, and model structural uncertainty. All sources of uncertainty are combined in the predictive uncertainty of the model [*Schoups and Vrugt*, 2010]. Uncertainty in simulated heads due to parameter uncertainty is represented by 95% credible intervals and total predictive uncertainty is represented by 95% prediction intervals. The 10,000 sampled parameter sets in the MCMC chains each contain an estimate of the seven parameters of the TFN model. Heads are simulated with each of the parameter sets (equation (2)). The 2.5 and 97.5 percentiles of the simulated heads give the 95% credible intervals. Residuals are added to the series of simulated heads to construct the prediction intervals. For each series of heads, a series of residuals is drawn from a zero-mean

normal distribution with the standard deviation from the same parameter set as used for simulating the heads. The 2.5 and 97.5 percentiles of the simulated heads plus residuals give the 95% prediction intervals. The credible intervals and prediction intervals are estimated for the complete lengths of the series of observed heads irrespective of the period used for calibration.

The Nash-Sutcliffe coefficient [*Nash and Sutcliffe*, 1970] is used to quantify the model fit and is calculated as follows

$$NS = 1 - \frac{\sum [h_o(t_i) - h(t_i)]^2}{\sum [h_o(t_i) - \bar{h}_o(t_i)]^2}, \tag{11}$$

where h is simulated with the set of parameters that, out of all the sampled parameter sets in the MCMC chains, gives the largest log likelihood. This parameter set approaches the maximum likelihood solution and will in the following be referred to as the "maximum likelihood estimate." Nash-Sutcliffe coefficients are calculated for three sets of observed heads: the calibration period, the validation period (all heads not used for calibration), and the entire set of observations. It is noted that due to the long response times of groundwater systems, time series of observed heads are rarely long enough to split into independent calibration and validation data sets [*Konikow and Bredehoeft*, 1992]. Heads observed during the validation period are influenced by stresses during the calibration period (or vice versa when the calibration period is after the validation period).

5. Study Area

This research focuses on time series from the Netherlands, a deltaic region with a temperate maritime climate. Most precipitation infiltrates into the subsurface where a dense drainage system discharges precipitation surplus [*De Vries*, 2007]. The groundwater table is generally shallow, with exceptions for ice-pushed hills in the central and eastern parts of the country and uplifted areas in the southeastern part [*De Vries*, 2007]. Annual precipitation is ~800 mm/yr, fairly evenly distributed over the year [*KNMI*, 2016a]. Makkink reference evaporation is ~540 mm/yr on average, but is considerably higher in summer than in winter, with ~8 mm/month in January and ~90 mm/month in July (weather station "De Bilt" [*KNMI*, 2016b]).

6. Synthetic Data

In this section, the influence of the length of the calibration period and observation frequency on the uncertainty of the model results is studied using synthetic data. Synthetic data enable analyzing the influence of the different model parameters by varying them one by one. Furthermore, long series of heads can be generated with frequencies of one observation per day, which are not available in practice.

"Observed" heads are generated on a daily basis for a 30 year period from 1 January 1975 to 31 December 2004 with precipitation and evaporation data from weather station "De Bilt" in the Netherlands [*KNMI*, 2015]. Heads h are generated with equation (2) and residuals are added to these heads to create observed heads h_o (equation (1)). The residuals r are generated using the inverse of equation (7), where the innovations v are drawn from a zero-mean normal distribution. The correlated residuals represent the difference between observed and modeled heads, which is caused by errors in the input data, errors in the output data, and model structural errors.

Parameter values are used that are representative for deltaic regions such as the Netherlands. First a system referred to as the "standard system" is considered. The results for the standard system are compared to results for systems with shorter and longer response times and for systems with fast and slowly decaying noise. The response time is characterized by the $t_{0.9}$ value, the time at which 90% of the response has taken place. The decay time of the noise is characterized by the $\tau_{0.9}$ value, the time it takes for a residual to decay by 90%.

The model is calibrated to the first 5, 10, and 20 years and to the total period of 30 years of "observed" heads to study the influence of the length of the calibration period. Additionally, the observed heads in each of these four periods are sampled with time steps of 1, 15, 30, and 90 days to analyze the influence of

Table 2. True Parameter Values Synthetic Standard System

Parameter	Value
f	0.9
d (m)	10.0
A (day)	500.0
a (day)	100.0
n	1.0
α (day)	50.0
σ_r (m)	0.2

the observation frequency. The model is fitted with the same precipitation and evaporation as used for generating the heads h.

Evaluation of the results is based on credible and prediction intervals. The intervals are calculated for the total period of 30 years irrespective of the period used for calibration. Nash-Sutcliffe coefficients (equation (11)) are not presented for the synthetic data, because they vary little for the calibration periods and time steps considered. In general, they approach the values calculated by comparing the generated heads h, without noise, to the generated observed heads h_o, which include noise. It is noted that it is easier to obtain a good model fit for the synthetic heads than for measured heads, because the structure and distribution of the generated errors correspond exactly to the noise model and likelihood function.

6.1. Synthetic Data Standard System

The parameter values of the standard system are given in Table 2. The response time of the standard system is 230 days. Based on parameter estimates for series of measured heads, the value of the decay factor of the noise α is half the value of model parameter a, such that the decay time of the noise is 115 days (since $n = 1$). The block response function, which is the impulse response function integrated over 1 day with unit recharge, is shown in Figure 1 (red line). The Nash-Sutcliffe coefficient is $\sim$0.8 when the generated heads h, without noise, are compared to the generated observed heads h_o. As an example, the observed heads of the total period of 30 years, sampled with a time step of 15 days (red dots), are shown in Figure 2b together with the maximum likelihood estimate of the head (blue line), and the 95% prediction interval (gray). Recharge calculated with the maximum likelihood estimate of the evaporation factor is shown in Figure 2a; the red line in Figure 2a represents the 1 year moving average of the recharge. The credible and prediction intervals for the standard system are shown in Figure 3 for the four described calibration periods (L; dark blue, light blue, red, and orange), each with four frequencies (time step Δt; 4 bars per color, with different hatch marks). The intervals are averaged over the total period of 30 years. The filled parts of the bars are the credible intervals and the complete bars are the prediction intervals.

A calibration period of 5 years (dark blue bars) gives a very large parameter and predictive uncertainty, indicated by very large intervals, even with an observation frequency of one observation per day. Increasing the length of the calibration period to 10 years (light blue bars) gives a major decrease in the intervals. The intervals further decrease if calibration periods of 20 years (red bars) and 30 years (orange bars) are used, but the improvement is small, especially going from 20 to 30 years. The observation frequency influences the intervals for calibration periods of 5 and 10 years, but with calibration periods of 20 and 30 years the influence of the observation frequency is small.

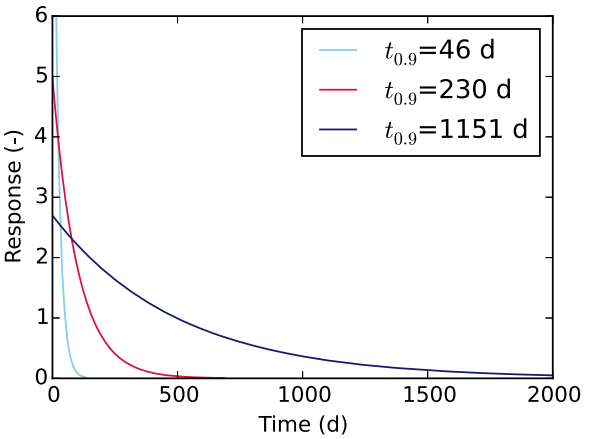


Figure 1. True block responses for synthetic systems with three different response times $t_{0.9}$ (different values of a).

Interestingly, the prediction intervals resulting from calibration on 5 years of daily head observations (first dark blue bar) are about 2 times as large as the prediction intervals obtained with 10 years of monthly observations (third light blue bar), while the total number of observations is 15 times as large for the first dark blue bar as compared to the third light blue bar. This indicates that the length of the calibration period is much more important than the total number of observations.

Furthermore, it can be seen that mainly the uncertainty due to parameter uncertainty (credible intervals) depends on the length of

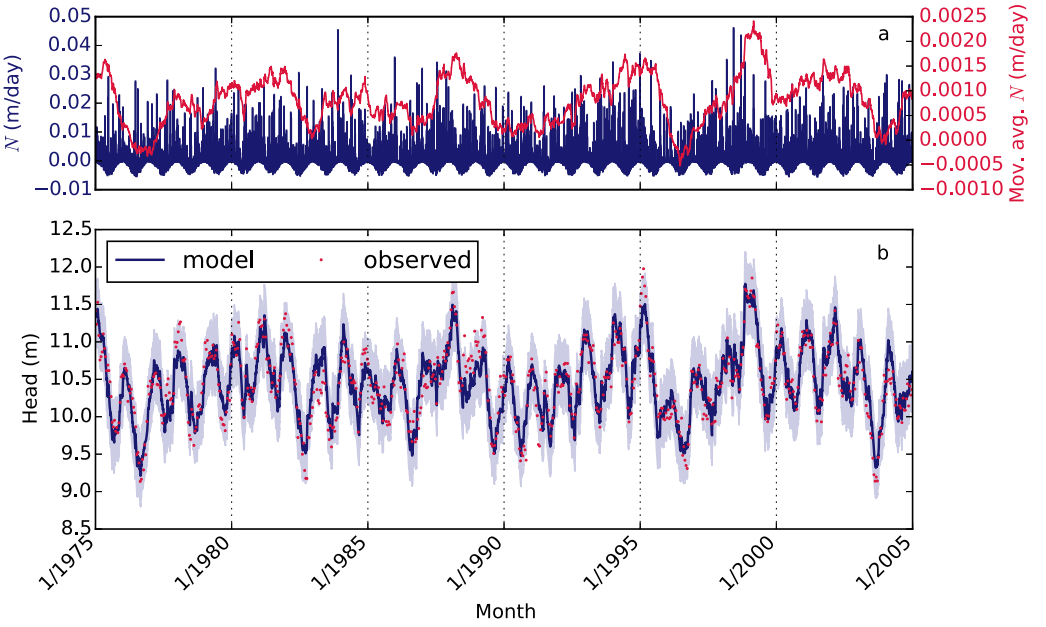


Figure 2. Synthetic standard system. Results for calibration period of 30 years with time step of 15 days. (a) Recharge calculated with maximum likelihood estimate of evaporation factor (blue bars) and 1 year moving average of recharge (red line). (b) Observed heads (red dots), maximum likelihood estimate of the head (blue line) with 95% prediction intervals (gray area).

the calibration period and observation frequency. When the parameter uncertainty approaches 0, the 95% prediction interval approaches 4 times the residual standard deviation ($\pm 2\sigma_r$), which in this case equals 0.8 m.

6.2. Synthetic Data Different Systems

The value of a is adjusted to 20 days and to 500 days to represent systems with a short response time of 46 days and a long response time of 1151 days (Figure 1). The parameter a influences the total fluctuation of the head as well. The value of A is adjusted to 260 and to 1350 days, respectively, such that the Nash-Sutcliffe coefficients of the generated h are ~0.8, similar to the standard system. In Figure 4, the credible and prediction intervals are shown for the system with the short response time, the standard system, and the system with the long response time. The intervals resulting from calibration on 5 years of head observations are relatively small for the system with a short response time in comparison to the other two systems, but they are still much larger than the intervals obtained with a calibration period of 10 years. Furthermore, the difference between the intervals of calibration periods of 10 and 20 years are largest for the system with the longest response time. However, the influence of the response time is relatively small considering the large difference in response times.

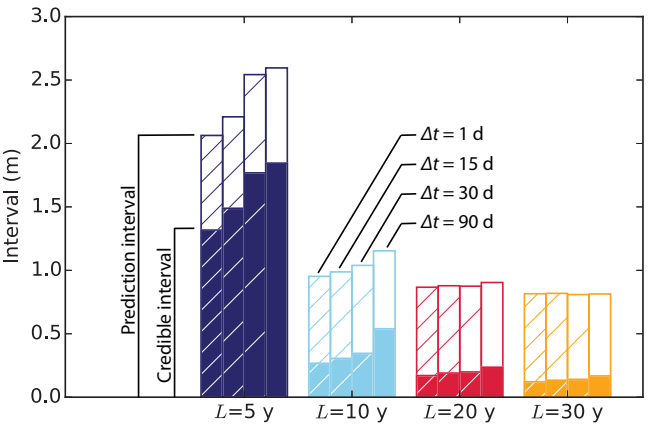


Figure 3. Average 95% credible intervals (filled parts of bars) and prediction intervals (complete bars) for total period of 30 years for standard system. Calibrated with periods of 5, 10, 20, and 30 years (colors) and time steps of 1, 15, 30, and 90 days (hatch marks).

The value of α is adjusted to 10 and 250 days to generate fast-decaying noise (decay time 23 days) and slowly decaying noise (decay time 576 days), respectively. The standard deviation of the residuals is kept constant. The residuals are shown in Figure 5. The response function is the same as for the standard system. The intervals of the systems with fast and slowly decaying noise are

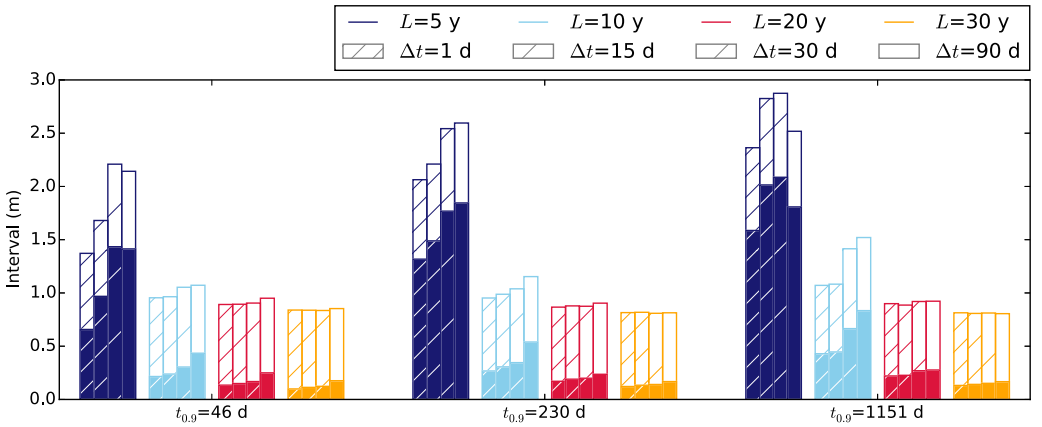


Figure 4. Average 95% credible intervals (filled parts of bars) and prediction intervals (complete bars) for total period of 30 years for models with three different response times. Standard system is shown in the middle. Calibrated with periods of 5, 10, 20, and 30 years and time steps of 1, 15, 30, and 90 days.

compared to those of the standard system in Figure 6. If noise decays very fast ($\tau_{0.9}$=23 days), calibration on 5 years of head observations already gives intervals similar to those obtained with longer calibration periods, unless the observation frequency is very small. For the system with slowly decaying noise ($\tau_{0.9}$=576 days), the interval still reduces when the length of the calibration period is increased from 20 to 30 years, while for the standard system this improvement is much smaller. The influence of the decay rate of the noise on the required length of the calibration period is significant: The slower the decay of the noise, the longer the calibration period that still gives a reduction in uncertainty.

Besides the response time and the decay rate of the noise, the residual standard deviation σ_r and response function parameter n were varied to 0.1 and 0.3 m, and to 0.5 and 1.5, respectively (results not shown). The residual standard deviation does not affect the required length of the calibration period and observation frequency, although a larger residual standard deviation results, obviously, in larger intervals. The impact of the variations of n was very small.

7. Measured Data

Eighteen time series of measured heads were selected from the national database on the subsurface in the Netherlands [*TNO*, 2015/2016]. Time series were selected that are at least 20 years long and for which the Nash-Sutcliffe coefficient of the maximum likelihood estimate obtained by calibration on the complete series of heads is at least 0.7. Series with a wide range of response times were selected, varying from $t_{0.9}$= 60 days to $t_{0.9}$ = 1179 days. A map with the locations of the 18 selected piezometers is shown in Figure 7. The closest weather stations with rainfall and evaporation data were used for each analysis. Details on the piezometers, mean water levels, geohydrological settings, and weather stations are given in Table S1 of Supporting Information.

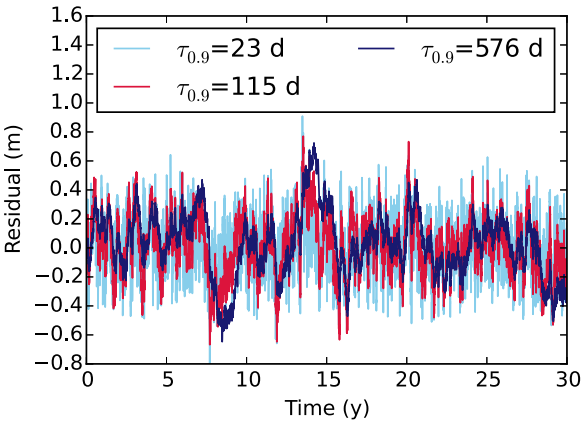


Figure 5. Generated residuals for systems with three different noise decay times $\tau_{0.9}$ (different values of α).

For each piezometer, the model is calibrated to the complete series and to periods of 5, 10, 20, and, if available, 30 years of observed heads. Results do not only depend on the length of the calibration period, but also on which part of a series is used for calibration. Therefore, the model is calibrated to a range of periods with starting dates at 5 year intervals. The heads were measured approximately 2 times per month. Time series with

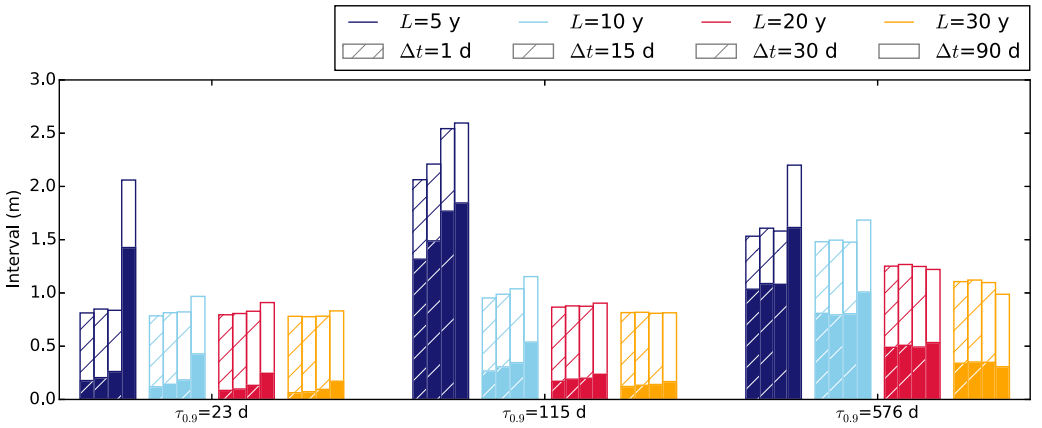


Figure 6. Average 95% credible intervals (filled parts of bars) and prediction intervals (complete bars) for total period of 30 years for models with three different noise decay times. Standard system is shown in the middle. Calibrated with periods of 5, 10, 20, and 30 years and time steps of 1, 15, 30, and 90 days.

frequencies of one observation per month and one observation per 3 months are approximated by sampling every second and sixth observed head, respectively, to analyze the influence of the observation frequency. In the following, first the results for one piezometer are discussed in detail, followed by the results for all piezometers.

7.1. Piezometer B32E0031

Piezometer B32E0031 is located in the central part of the Netherlands (number 14 in Figure 7). The piezometer is installed in an unconfined, sandy aquifer with a mean head that is 1.2 m below the surface; the screen is 19–20 m below the surface. The series of observed heads runs from November 1968 to October 2003. The head is modeled with precipitation and evaporation data from weather stations Putten and De Bilt, respectively. The observed data and model results for calibration on the complete series are shown in Figure 8. The Nash-Sutcliffe coefficient is relatively high (0.83), but the mean prediction interval is 0.80 m, which is still 45% of the difference between the 2.5 and 97.5 percentiles of the observed heads. In the following, the difference between the 2.5 and 97.5 percentiles of the observed heads will be referred to as the "head range." The estimated response time $t_{0.9}$ is 559 days.

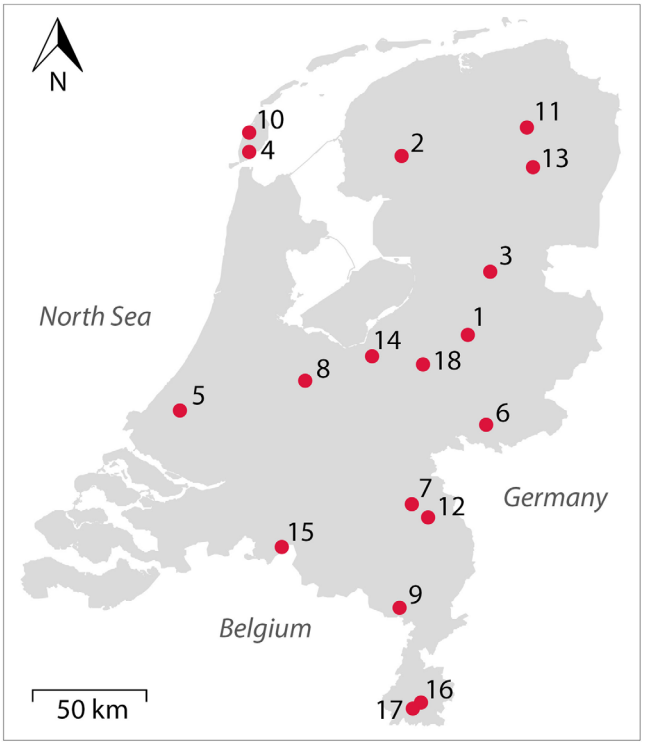


Figure 7. Locations of piezometers in the Netherlands. See Table S1 in Supporting Information for details.

Intervals and Nash-Sutcliffe coefficients obtained with calibration periods of different lengths and starting times are given in Figure 9. The intervals are normalized through division by the head range, to enable comparison to the results of other piezometers in the next section. The filled parts of the bars in Figure 9a represent the credible intervals and the total bars the prediction intervals. The lengths of the calibration periods are indicated with colors. The credible intervals and prediction intervals are summarized on the right-hand

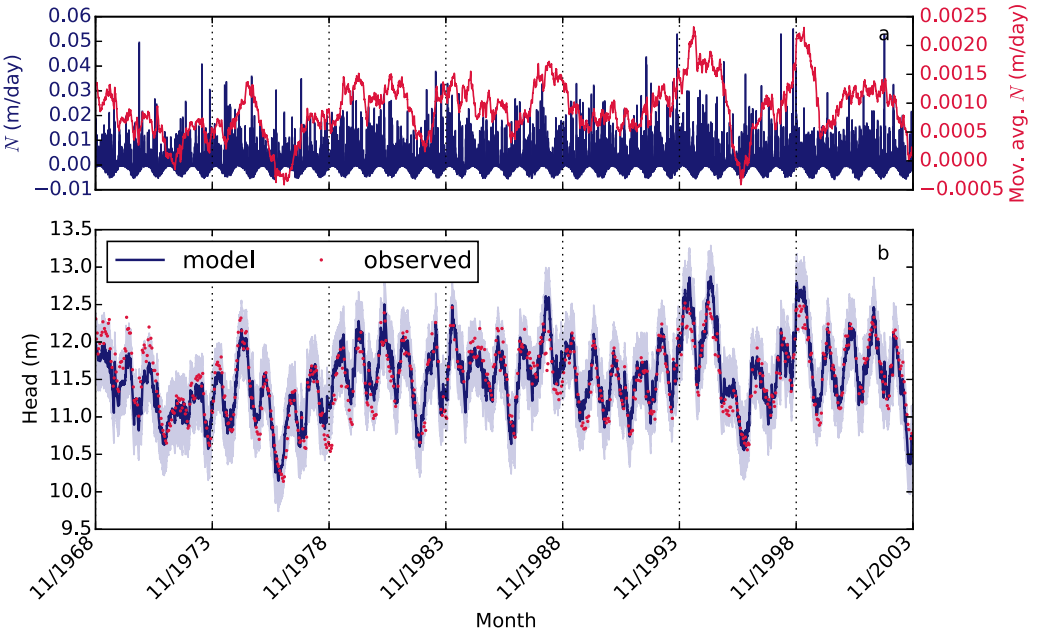


Figure 8. Recharge, observed, and modeled heads for piezometer B32E0031. Results obtained by calibration on all observed heads. (a) Recharge calculated with maximum likelihood estimate of evaporation factor (blue bars) and 1 year moving average of recharge (red line). (b) Observed heads (red dots), maximum likelihood estimate of the head (blue line) with 95% prediction intervals (gray area).

side of the figure. The dots indicate the averages found with calibration periods of a certain length, and the colored vertical lines range from the smallest to the largest values. Similar to the synthetic data (section 6), the parameter uncertainty (credible intervals) decreases when the length of the calibration period is increased,

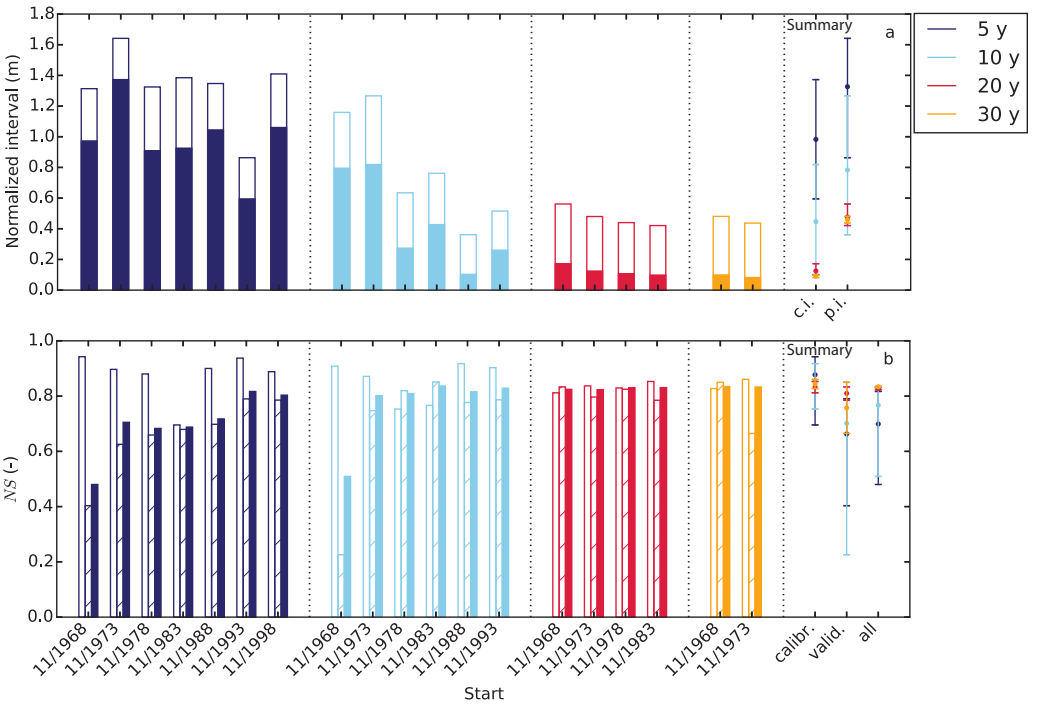


Figure 9. Results for piezometer B32E0031, obtained with calibration periods of 5, 10, 20, and 30 years starting at seven dates, with frequency of approximately two observations per month. (a) Normalized 95% credible intervals (filled parts of bars) and prediction intervals (complete bars) averaged over the complete period. (b) Nash-Sutcliffe coefficients for calibration period (empty bars), validation period (hatched bars), and for complete series (filled bars). Right-hand sides: summaries. Colored vertical lines range from minimum to maximum values, indicated by horizontal line markers. Dots indicate averages.

and the prediction interval eventually approaches 4 times the estimated residual standard deviation. Calibration on 5 years of observed heads results in a large uncertainty, as shown in Figure 9a (dark blue). For six of the seven calibration periods of 5 years, the 95% prediction intervals are at least 30% larger than the head range. Calibration on 10 years of observations gives a large improvement for some starting dates, but hardly any improvement for other starting dates. It is noted that for this specific example the largest intervals are found for calibration periods at the beginning of the measurements, but this seems to be a coincidence and varies for the other piezometers considered. Overall, relatively small intervals are obtained with 20 years of observations. The intervals do not decrease much further when the length of the calibration period is increased to 30 years. The smallest 95% prediction intervals are ~40–50% of the head range.

The Nash-Sutcliffe coefficients do not improve anymore either when extending the calibration period from 20 to 30 years, as can be seen in Figure 9b. The Nash-Sutcliffe coefficients calculated with the subset of observed heads used for calibration are represented by empty bars, the Nash-Sutcliffe coefficients calculated with the validation data are represented by hatched bars, and the Nash-Sutcliffe coefficients calculated with the complete series of observed heads are represented by filled bars. The Nash-Sutcliffe coefficients are summarized on the right-hand side of the figure. For a calibration period of 5 years, the Nash-Sutcliffe coefficients calculated for the calibration data are above 0.8 for six of the seven periods, but the Nash-Sutcliffe coefficients calculated for the validation data and for the complete series are generally lower. For calibration periods of 10 years, the differences between the Nash-Sutcliffe coefficients calculated with the calibration data, validation data, and total period are smaller, except for the calibration period starting in 1968. The Nash-Sutcliffe coefficients of the validation period and those of all the data are similar to the Nash-Sutcliffe coefficients of the calibration data for calibration periods of 20 years. Further increasing the calibration period to 30 years leads to a decrease in the Nash-Sutcliffe coefficient for the validation data for the calibration period starting in 1973, but it is noted that the calibration periods of 30 years approach the total period, such that the number of observations left for validation is very small.

7.2. Analysis of All 18 Piezometers

7.2.1. Influence Length of Calibration Period on Uncertainty Intervals

For each piezometer, the credible and prediction intervals are estimated for different lengths and starting times of the calibration periods. The intervals are normalized through division by the head range in each piezometer. The intervals for calibration periods of 10, 20, and, if available, 30 years are plotted against the maximum likelihood estimates of the response times of the 18 systems, obtained by calibration on the complete series (Figure 10). Three colored lines are plotted for each piezometer, similar to the summaries for piezometer B32E0031 on the right-hand side of Figure 9a. The intervals for calibration periods of 5 years are left out of Figure 10, because they vary widely and are often very large.

In general, the credible intervals (Figure 10a), indicating parameter uncertainty, decrease if the length of the calibration period is increased. However, the prediction intervals (Figure 10b), indicating total predictive uncertainty, are sometimes a little smaller for shorter calibration periods. This can be seen from the blue vertical lines in Figure 10b, which extend below the red and orange lines for the majority of the piezometers. The difference between the credible and prediction intervals is determined by the residuals. Results indicate that on average the standard deviation of the residuals is underestimated for shorter calibration periods, explaining why the prediction intervals are sometimes smaller. A figure with the maximum likelihood estimates of the standard deviation of the residuals for the different calibration periods is included in Figure S2 of Supporting Information.

The prediction intervals obtained with a calibration period of 5 years are on average 1.22 times the head range, and eight of the piezometers even have one or more calibration periods that give prediction intervals larger than 2 times the head range. On average the prediction intervals are reduced to 0.63 times the head range if the length of the calibration period is increased to 10 years (Figure 10b). However, the blue vertical lines show that there is a large spread in these intervals for most piezometers, so the uncertainty depends on which period of 10 years is used. The prediction intervals obtained with calibration periods of 20 years (red lines) and 30 years (orange lines) vary much less and are on average 0.53 and 0.54 times the head range. Concluding, accurate results cannot be expected with a calibration period of 5 years, relatively accurate results are generally obtained with calibration periods of 20 years, while further improvement by

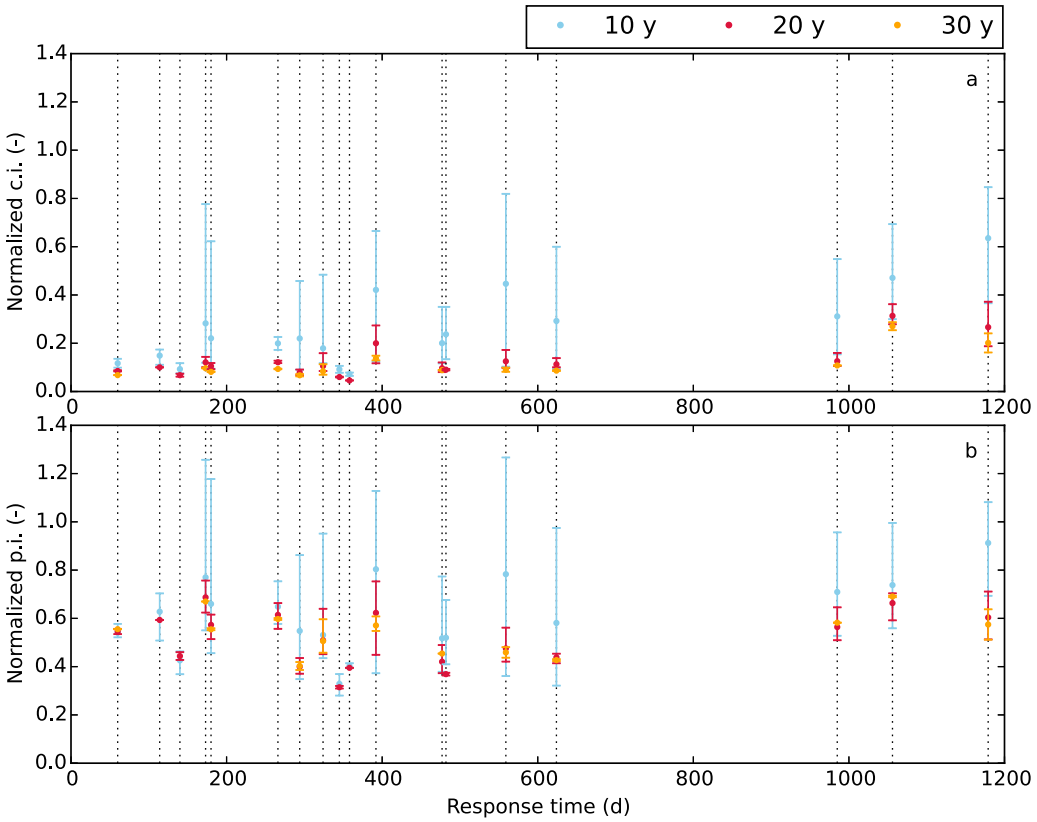


Figure 10. Normalized 95% (a) credible intervals and (b) prediction intervals for calibration periods of 10, 20, and, if available, 30 years, against estimated response time for each location. Colored vertical lines range from minimums to maximums, indicated by horizontal line markers. Dots indicate averages.

calibration on 30 years of observations is small to negligible. It is noted that for five of the piezometers no calibration periods of 30 years are available, because the total period of observed heads is too short.

The influence of the length of the calibration period is not clearly related to the response time. The estimated response times vary between 60 and 1200 days. For the three piezometers with the shortest response times, the credible intervals obtained with calibration periods of 10 years (blue lines) seem to be relatively small for all calibration periods, but this also holds for some of the piezometers with larger response times. In general, the length of the calibration period that results in relatively small intervals is much larger than the response time.

Although no relation between the response time and the required length of the calibration period was observed, the required length of the calibration period seems to be related to the decay time of the noise, similar to the results for the synthetic data (section 6.2). In Figure 11 the credible intervals are plotted against the maximum likelihood estimates of the decay time of the noise for each specific calibration period, for all piezometers. The intervals obtained with calibration periods of 10, 20, and 30 years are shown as blue, red, and orange dots, respectively. The credible intervals generally increase with the decay time of the noise. This trend is not as clearly seen when the prediction intervals are plotted against the decay time of the noise (not shown).

It can be expected that the decay time of the noise is related to the response time of the system. Therefore, it may be surprising that a relationship was observed between the required length of the calibration period and the decay time of the noise $\tau_{0.9}$ and not between the required length of the calibration period and the response time $t_{0.9}$. The correlation coefficient of the maximum likelihood estimates of $\tau_{0.9}$ and $t_{0.9}$ is 0.79 when calibrating on the complete time series, but much lower when calibrating on shorter sections of the time series. Differences in the estimated decay time of the noise for different parts of the series might be one of the causes for the large variation in the intervals for calibration periods of 5 and 10 years.

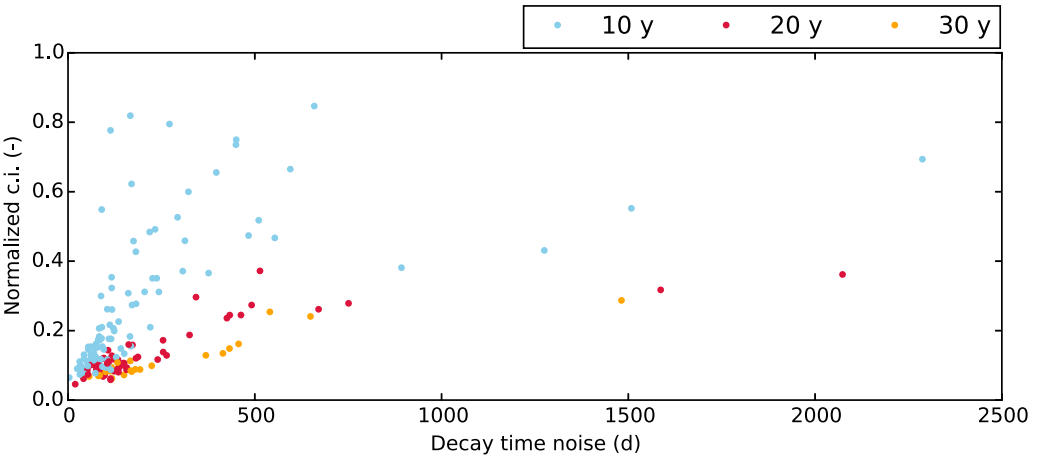


Figure 11. Normalized 95% credible intervals for calibration periods of 10, 20, and 30 years versus estimated decay time of the noise for each calibration period.

7.2.2. Influence Length of Calibration Period on Nash-Sutcliffe Coefficients

The Nash-Sutcliffe coefficients are calculated for calibration periods of 5, 10, 20, and 30 years. In this section, the Nash-Sutcliffe coefficients are reported for the complete series of observed heads h_o (equation (11)). The calibration periods of 20 and, especially, 30 years approach the complete periods of observed heads for many of the series, such that the number of observations left is too small for an independent validation. Nash-Sutcliffe coefficients for the validation period are likely smaller (see, e.g., Figure 9b).

The Nash-Sutcliffe coefficients found with calibration periods of 10, 20 and, if available, 30 years are plotted against the estimated response time for each piezometer in Figure 12, similar to the summary on the right-hand side of Figure 9b for piezometer B32E0031. Calibration periods of 5 years (not included in the figure) may result in high or low Nash-Sutcliffe coefficients, depending on which period of 5 years is used for calibration and vary between −0.44 and 0.93, with an average of 0.66. The average values of the Nash-Sutcliffe coefficient for each piezometer, indicated by the dots in Figure 12, increase if the length of the calibration period is increased (the red and orange dots are above the blue dots). The average Nash-Sutcliffe coefficient for all piezometers for a calibration period of 10 years (blue) is 0.72. For a calibration period of 20 years (red) it is 0.76 and for a calibration period of 30 years (orange) it is 0.77. Furthermore, the vertical lines in the figure show that the spread for the calibration periods of 10 years (blue) is larger than that for calibration periods of 20 years (red) and 30 years (orange). Moreover, some very low coefficients are obtained with calibration periods of 10 years. The smallest Nash-Sutcliffe coefficient for a calibration period of 10 years is

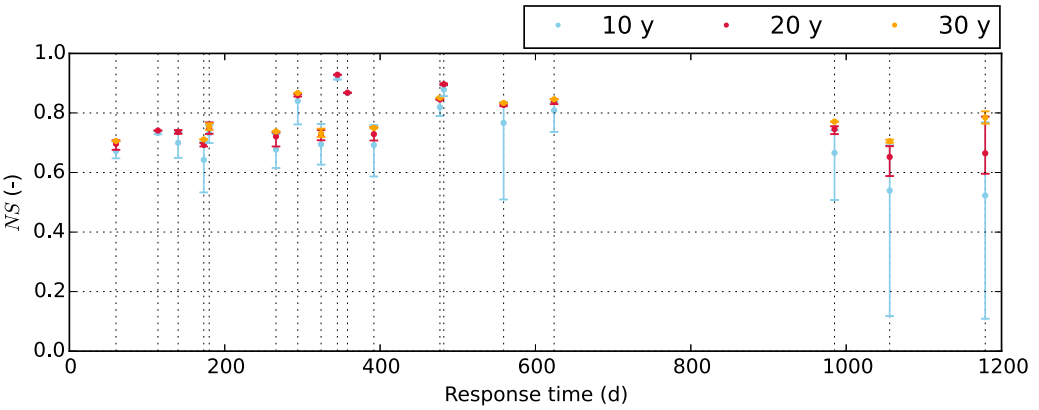


Figure 12. Nash-Sutcliffe coefficients for calibration periods of 10, 20 and, if available, 30 years, against estimated response time for each location. Nash-Sutcliffe coefficients are calculated using complete series of heads. Colored vertical lines range from minimum to maximum values, indicated by horizontal line markers. Dots indicate averages.

Table 3. Influence Observation Frequency[a]

		Credible Intervals		Prediction Intervals		Nash-Sutcliffe	
		1 Month	3 Months	1 Month	3 Months	1 Month	3 Months
5 years	avg.	1.5	4.1	1.2	2.2	−0.01	−0.19
	max.	7.1	53.8	3.5	20.3	−0.79	−10.00
10 years	avg.	1.4	2.5	1.1	1.4	0.00	−0.01
	max.	4.1	8.1	2.3	3.2	−0.29	−1.23
20 years	avg.	1.1	1.7	1.0	1.1	0.01	0.01
	max.	1.5	4.5	1.2	1.8	−0.02	−0.03
30 years	avg.	1.1	1.4	1.0	1.0	0.01	0.01
	max.	1.3	2.1	1.0	1.2	−0.01	−0.02

[a]Intervals obtained by calibration on heads observed with frequencies of 1 per month and 1 per 3 months divided by intervals obtained by calibration on heads observed with a frequency of 2 per month. Average relative intervals (–) and maximum relative intervals (–) are given. Nash-Sutcliffe coefficients (–) obtained by calibration on heads observed with frequencies of 1 per month and 1 per 3 months minus Nash-Sutcliffe coefficients obtained by calibration on heads observed with a frequency of 2 per month. Nash-Sutcliffe coefficients were calculated with the complete series of observed heads. Average differences and largest decreases in Nash-Sutcliffe coefficients are given.

0.11. For a calibration period of 20 years the smallest value is 0.59 and for a calibration period of 30 years it is 0.70.

The influence of the calibration period on the Nash-Sutcliffe coefficients does not seem to be strongly related to the response time, although the largest variation in the Nash-Sutcliffe coefficients is found for the two piezometers with the longest response times. Also, the smallest Nash-Sutcliffe coefficients for calibration periods of 10 and 20 years are found for these two piezometers. No relationship between the influence of the calibration period on the Nash-Sutcliffe coefficients and the decay time of the noise was found.

7.2.3. Influence Observation Frequency

The influence of the observation frequency is analyzed by computing the credible and prediction intervals for frequencies of one observation per month and one observation per 3 months and dividing them by the intervals obtained with a frequency of two observations per month. Additionally, the differences between the Nash-Sutcliffe coefficients for the different frequencies are calculated. Results are presented in Table 3.

With a calibration period of 10 years, the credible interval is 1.4 times as large on average if the observation frequency is decreased from two observations per month to one observation per month, while the prediction interval is only 1.1 times as large on average. However, the maximum differences are large. The maximum increases in the credible and prediction intervals are 4.1 and 2.3, respectively, when the frequency is decreased from two observations per month to one observation per month. The average differences in the Nash-Sutcliffe coefficients found with the different observation frequencies are negligible if the length of the calibration period is 10 years. The largest decrease in the Nash-Sutcliffe coefficients is −0.29 if the frequency is decreased from two observations per month to one observation per month, and −1.23 if the frequency is decreased from two observations per month to one observation per 3 months.

Changes in the uncertainty intervals and Nash-Sutcliffe coefficients due to changes in the observation frequency are much smaller for longer calibration periods. With a calibration period of 20 years, the prediction intervals obtained with frequencies of one and two observations per month are on average the same. The largest increase is 1.2. The credible intervals are 1.1 times as large on average, while the largest increase is 1.5. The differences are larger for a frequency of one observation per 3 months. Surprisingly, on average the Nash-Sutcliffe coefficients increase slightly if the observation frequency is decreased. For some of the individual cases the Nash-Sutcliffe coefficients decrease, but even if the frequency is decreased to one observation per 3 months the largest decrease is small, with −0.03.

For the series considered, the longer the calibration period, the smaller the influence of the observation frequency. Furthermore, the observation frequency has a larger effect on the parameter uncertainty (the credible intervals) than on the prediction intervals.

The length of the calibration period seems to be more important than the observation frequency. The average normalized credible and prediction intervals estimated with a calibration period of 10 years and frequency of two observations per month are 0.28 and 0.63, respectively (Table 4). If the period is doubled to 20 years and the frequency is halved to one observation per month, meaning that the total numbers of

Table 4. Average Normalized Credible and Prediction Intervals (–) and Average Nash-Sutcliffe Coefficients (–), for Calibration Periods of 10 Years With Two Observations per Month and Calibration Periods of 20 Years With One Observation per Month and One Observation per 3 Months

	c.i.	p.i.	N.S.
10 years, 2 per month	0.28	0.63	0.72
20 years, 1 per month	0.15	0.54	0.77
20 years, 1 per 3 months	0.21	0.59	0.78

observations are approximately the same, the average normalized credible and prediction intervals are reduced to 0.15 and 0.54, respectively. Even with a frequency of one observation per 3 months for 20 years, the intervals are still smaller than the intervals obtained with two observations per month for 10 years (on average 0.21 and 0.59, respectively). Moreover, average Nash-Sutcliffe coefficients calculated for the complete time series increase from 0.72 to 0.77 if the length of the calibration period is increased from 10 to 20 years, while the frequency is decreased from two observations per month to one observation per month.

8. Discussion

This study is limited to time series of head observations that could be simulated well with a linear TFN model with precipitation and evaporation as the only input series. Other series may require more stresses, for example pumping [e.g., *Obergfell et al*., 2013; *Shapoori et al*., 2015a; *Von Asmuth et al*., 2008], which may influence the required length of the calibration period and observation frequency. Changes to a system may cause heterogeneity in observed series of heads. The longer the series, the more likely it is that this problem occurs. Series with significant heterogeneity, such as a linear trend or step changes, were not used for this study.

A good fit was obtained for the series considered in this paper with a linear TFN model. The actual response is, most likely, at least slightly nonlinear. In deltaic regions with shallow groundwater tables and a temperate maritime climate, nonlinearity may be caused by, e.g., variations in moisture content of the unsaturated zone, variations in soil layers, and the presence of drainage systems [e.g., *Knotters*, 2001]. Nonlinearity was not included in the synthetic data, resulting in a model fit that was much less dependent on the length of the calibration period than for the measured data. Inclusion of the nonlinear response in the time series model may reduce the required length of the calibration period for the measured data, but this was not studied. Nonlinear behavior may also explain why residual standard deviations are underestimated for short calibration periods.

In the study area, runoff does not play a major role, as rainfall rarely exceeds infiltration capacity [*De Vries*, 2007]. However, relatively wet and relatively dry years may be identified in the weather data, resulting in relatively high peaks and low valleys in the head data (e.g., Figure 8). *Knotters and Van Walsum* [1997] and *Knotters* [2001] stated that a proper calibration period should include head variations that reflect the entire fluctuation range. Calibration on a 5 year period that includes a very wet and a very dry year (i.e., 1993–1998 in Figure 8) resulted in relatively small credible and prediction intervals (sixth dark blue bar in Figure 9a), but still significantly larger than the credible and prediction intervals using 20 years of calibration data (red bars in Figure 9a). Climates with long wet and dry periods and significant runoff, e.g., (sub)tropical monsoon climates, can likely not be simulated accurately with a linear model, but require a nonlinear model [e.g., *Peterson and Western*, 2014]. It is expected that longer calibration periods are needed for climates with long wet and dry periods, but this was not studied.

Sensitivity to the calibration data may depend on the calibration method used [*Schoups and Vrugt*, 2010; *Sorooshian et al*., 1983]. In this research the MCMC method was used and it was assumed that the innovations have a normal distribution. Research is needed to find out how deviations from normality impact the influence of the length of the calibration period and observation frequency on the uncertainty. Furthermore, wide uniform prior distributions were used to represent uninformative priors and focus on the information contained in the data. In practice, if prior knowledge is available, narrower or other distributions may be used, which may limit the uncertainty of the model results.

Results indicate that the required length of the calibration period is related to the decay time of the noise. The larger the temporal correlation in the noise, the more difficult it is to distinguish the deterministic and stochastic parts of the model, and the longer the required calibration period. For long calibration periods, strong correlation was found between the decay time of the noise and the response time of the system.

One of the causes of this correlation may be that an error in observed precipitation or evaporation has a longer effect on the model results when the response time is longer. A long decay time may also indicate that the relationship between recharge and heads is not sufficiently described by the (linear) transfer part of the model. Although only time series were used for which the linear model gives a good fit (Nash-Sutcliffe coefficients of at least 0.7), it cannot be excluded that a better model may reveal a relationship between the required calibration period and the response time.

9. Conclusions

The influence of the length of the calibration period and observation frequency on predictive uncertainty in time series modeling of groundwater dynamics was investigated. Head series were used that could be modeled with a linear Transfer Function-Noise (TFN) model. Input series were precipitation and evaporation in the Netherlands, a deltaic region with predominantly shallow groundwater tables in a temperate maritime climate. Uncertainty in model predictions, caused by data uncertainty, model structural uncertainty, and parameter uncertainty, is inevitable, even if large amounts of calibration data are available. Increasing the length of the calibration period and observation frequency decreases the uncertainty of the estimated parameter values. When the parameter uncertainty approaches 0, the 95% prediction intervals approach 4 times the standard deviation of the residuals ($\pm 2\sigma_r$).

Results for both synthetic and observed data indicate that the length of the calibration period is much more important than the total number of observations. Long calibration periods with a small observation frequency contain more information than short calibration periods with a high observation frequency, while the total number of observations is similar. This means that a short calibration period cannot be compensated by a high observation frequency.

No strong relation between the required length of the calibration period and the response time was observed, but results indicate that the required length of the calibration period is related to the decay time of the noise. The decay of the noise is slower when the temporal correlation in the noise is larger. As a result, it is more difficult to distinguish the deterministic and stochastic parts of the model, and the required calibration period is longer. The estimated response time of the system and the decay time of the noise are strongly correlated when calibrating on long series of observed heads, but correlation is much weaker when calibrating on shorter sections of the series. Further research is needed to study how noise decay is related to the characteristics of the system.

Relatively long calibration periods are required to minimize the uncertainty, both for systems with short and long response times. For the 18 piezometers studied, the 95% prediction intervals obtained with a calibration period of 5 years were on average ~20% larger than the head range. Only for cases with very fast decaying noise, relatively certain results were obtained with a calibration period of 5 years. The results improved significantly for calibration periods of 10 years, but varied widely depending on the specific period used. For calibration periods of 20 years variation was much smaller. The credible and prediction intervals for the 18 piezometers could be reduced to ~10% and ~50% of the head range, respectively. Further improvement was relatively small when the calibration period was increased to 30 years.

Appendix A: Innovation Standard Deviation

The time-step-dependent standard deviation of the innovations $\sigma_v(\Delta t)$ can be written as a function of the standard deviation of the residuals σ_r. An equation for the residuals is obtained from equation (7) as

$$r(t)=e^{-\Delta t/\alpha}r(t-\Delta t)+v(t). \tag{A1}$$

Taking the variance of both sides of equation (A1) gives

$$\sigma_r^2(t)=e^{-2\Delta t/\alpha}\sigma_r^2(t-\Delta t)+\sigma_v^2(t)+2e^{-\Delta t/\alpha}\mathrm{Cov}(r(t-\Delta t),v(t)). \tag{A2}$$

Assuming that the residual standard deviation is time-independent and that the covariance between $r(t-\Delta t)$ and $v(t)$ is zero, the innovation and residual standard deviations are related as

$$\sigma_v(\Delta t)=\sigma_r\sqrt{1-e^{-2\Delta t/\alpha}}. \quad \text{(A3)}$$

Appendix B: Reparameterizations

The averages are subtracted from the precipitation P and reference evaporation E to reduce problems with mixing of the MCMC chains due to correlation between the evaporation factor f and the base elevation d. Instead of the distribution of d, the distribution of d^* [L] is estimated

$$h(t)=d^*+\int_{-\infty}^{t}(P(\tau)-\bar{P})\theta(t-\tau)d\tau-f\int_{-\infty}^{t}(E(\tau)-\bar{E})\theta(t-\tau)d\tau. \quad \text{(B1)}$$

The trace of d^* is transformed back to d as

$$d=d^*-A(\bar{P}-f\bar{E}). \quad \text{(B2)}$$

The parameter σ^* [L] is introduced to reduce mixing problems due to correlation between the decay factor of the noise α and the standard deviation of the residuals σ_r.

$$\sigma^*=\sqrt{1-e^{-2\overline{\Delta t}/\alpha}}\sigma_r, \quad \text{(B3)}$$

where $\overline{\Delta t}$ is the average time step between the observations. The innovation standard deviation (equation (8)) may be written as

$$\sigma_v(\Delta t)=\frac{\sqrt{1-e^{-2\Delta t/\alpha}}}{\sqrt{1-e^{-2\overline{\Delta t}/\alpha}}}\sigma^*. \quad \text{(B4)}$$

The trace of σ^* is transformed back to σ_r as

$$\sigma_r=\frac{1}{\sqrt{1-e^{-2\overline{\Delta t}/\alpha}}}\sigma^*. \quad \text{(B5)}$$

Acknowledgments
Python scripts to reproduce Figures 1−6 and 8−12, including all data, are available from the authors.

References

Beck, M. B. (1987), Water quality modeling: A review of the analysis of uncertainty, *Water Resour. Res.*, *23*(8), 1393–1442, doi:10.1029/WR023i008p01393.
Berendrecht, W. L., A. W. Heemink, F. C. Van Geer, and J. C. Gehrels (2003), Decoupling of modeling and measuring interval in groundwater time series analysis based on response characteristics, *J. Hydrol.*, *278*(1–4), 1–16, doi:10.1016/S0022-1694(03)00075-1.
Box, G. E. P., and G. M. Jenkins (1970), *Time Series Analysis, Forecasting and Control*, Holden-Day, San Francisco, Calif.
Box, G. E. P., and G. C. Tiao (1992), *Bayesian Inference in Statistical Analysis*, John Wiley, New York.
De Vries, J. J. (2007), Groundwater, in *Geology of the Netherlands*, edited by T. E. Wong, D. A. J. Batjes, and J. de Jager, pp. 295–315, R. Netherlands Acad. of Arts and Sci., Amsterdam.
Dunn, W. L., and J. K. Shultis (2012), Markov chain Monte Carlo, in *Exploring Monte Carlo Methods*, pp. 133–169, Elsevier, Amsterdam, doi:10.1016/B978-0-444-51575-9.00006-3.
Finke, P. A., D. J. Brus, M. F. P. Bierkens, T. Hoogland, M. Knotters, and F. de Vries (2004), Mapping groundwater dynamics using multiple sources of exhaustive high resolution data, *Geoderma*, *123*, 23–39, doi:10.1016/j.geoderma.2004.01.025.
Fu, J., and J. J. Gómez-Hernández (2009), Uncertainty assessment and data worth in groundwater flow and mass transport modeling using a blocking Markov chain Monte Carlo method, *J. Hydrol.*, *364*, 328–341, doi:10.1016/j.jhydrol.2008.11.014.
Hassan, A. E., H. M. Bekhit, and J. B. Chapman (2009) Using Markov chain Monte Carlo to quantify parameter uncertainty and its effect on predictions of a groundwater flow model, *Environ. Modell. Software*, *24*, 749–763, doi:10.1016/j.envsoft.2008.11.002.
Juston, J., J. Seibert, and P. O. Johansson (2009), Temporal sampling strategies and uncertainty in calibrating a conceptual hydrological model for a small boreal catchment, *Hydrol. Processes*, *23*, 3093–3109, doi:10.1002/hyp.7421.
KNMI (2015), Daggegevens van het weer in Nederland. [Available at http://projects.knmi.nl/klimatologie/daggegevens/selectie.cgi.]
KNMI (2016a), Zware Neerslag. [Available at https://www.knmi.nl/kennis-en-datacentrum/uitleg/zware-neerslag.]
KNMI (2016b), Verdamping in Nederland. [Available at https://www.knmi.nl/kennis-en-datacentrum/achtergrond/verdamping-in-nederland]
Knotters, M. (2001), Regionalised time series models for water table depths, PhD thesis, Wageningen Univ., Wageningen, Netherlands. [Available at http://library.wur.nl/WebQuery/wurpubs/fulltext/121258.]
Knotters, M., and P. E. V. Van Walsum (1997), Estimating fluctuation quantities from time series of water-table depths using models with a stochastic component, *J. Hydrol.*, *197*, 25–46, doi:10.1016/S0022-1694(96)03278-7.
Konikow, L. F., and J. D. Bredehoeft (1992) Ground-water models cannot be validated, *Adv. Water Resour.*, *15*, 75–83, doi:10.1016/0309-1708(92)90033-X.
Kuczera, G., and E. Parent (1998), Monte Carlo assessment of parameter uncertainty in conceptual catchment models: The Metropolis algorithm, *J. Hydrol.*, *211*, 69–85, doi:10.1016/S0022-1694(98)00198-X.
Liu, Y., and H. V. Gupta (2007), Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, *Water Resour. Res.*, *43*, W07401, doi:10.1029/2006WR005756.

Laloy, E., B. Rogiers, J. A. Vrugt, D. Mallants, and D. Jacques (2013), Efficient posterior exploration of a high-dimensional groundwater model from two-stage Markov chain Monte Carlo simulation and polynomial chaos expansion, *Water Resour. Res.*, *49*, 2664–2682, doi:10.1002/wrcr.20226.

Lu, D., M. Ye, and M. C. Hill (2012), Analysis of regression confidence intervals and Bayesian credible intervals for uncertainty quantification, *Water Resour. Res.*, *48*, W09521, doi:10.1029/2011WR011289.

Marshall, L., D. Nott, and A. Sharma (2004), A comparative study of Markov chain Monte Carlo methods for conceptual rainfall-runoff modeling, *Water Resour. Res.*, *40*, W02501, doi:10.1029/2003WR002378.

McMillan, H., B. Jackson., M. Clark, D. Kavetski, and R. Woods (2011), Rainfall uncertainty in hydrological modelling: An evaluation of multiplicative error models, *J. Hydrol.*, *400*(1–2), 83–94, doi:10.1016/j.jhydrol.2011.01.026.

Moradkhani, H., and S. Sorooshian (2009), General review of rainfall-runoff modeling: Model calibration, data assimilation, and uncertainty analysis, in *Hydrological Modelling and the Water Cycle: Coupling the Atmospheric and Hydrological Models*, edited by S. Sorooshian et al., pp. 1–24, Springer Science & Business Media, Netherlands, doi:10.1007/978-3-540-77843-1_1.

Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models part I—A discussion of principles, *J. Hydrol.*, *10*, 282–290, doi:10.1016/0022-1694(70)90255-6.

Obergfell, C., M. Bakker, W. J. Zaadnoordijk, and K. Maas (2013), Deriving hydrogeological parameters through time series analysis of groundwater head fluctuations around well fields, *Hydrogeol. J.*, *21*, 987–999, doi:10.1007/s10040-013-0973-4.

Patil, A., D. Huard, and C. J. Fonnesbeck (2010), PyMC: Bayesian stochastic modelling in Python, *J. Stat. Software*, *35*(4), 1–81, doi:10.18637/jss.v035.i04.

Perrin, C., L. Oudin, V. Andreassian, C. Rojas-Serna, C. Michel, and T. Mathevet (2007), Impact of limited streamflow data on the efficiency and the parameters of rainfall-runoff models, *Hydrol. Sci. J.*, *52*(1), 131–151, doi:10.1623/hysj.52.1.131.

Peterson, T. J., and A. W. Western (2014), Nonlinear time-series modeling of unconfined groundwater head, *Water Resour. Res.*, *50*, 8330–8355, doi:10.1002/2013WR014800.

Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks (2010), Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, *Water Resour. Res.*, *46*, W05521, doi:10.1029/2009WR008328.

Schoups, G., and J. A. Vrugt (2010), A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, *Water Resour. Res.*, *46*, W10531, doi:10.1029/2009WR008933.

Shapoori, V., T. J. Peterson, W. Western, and J. F. Costelloe (2015a), Top-down groundwater hydrograph time-series modeling for climate-pumping decomposition, *Hydrogeol. J.*, *23*, 819–836, doi:10.1007/s10040-014-1223-0.

Shapoori, V., T. J. Peterson, W. Western, and J. F. Costelloe (2015b), Decomposing groundwater head variations into meteorological and pumping components: A synthetic study, *Hydrogeol. J.*, *23*, 1431–1448, doi:10.1007/s10040-015-1269-7.

Sorooshian, S., V. K. Gupta, and J. L. Fulton (1983), Evaluation of maximum likelihood parameter estimation techniques for conceptual rainfall-runoff models: Influence of calibration data variability and length on model credibility, *Water Resour. Res.*, *19*(1), 251–259, doi:10.1029/WR019i001p00251.

Thyer, M., and G. Kuczera (2003) A hidden Markov model for modelling long-term persistence in multi-site rainfall time series 1. Model calibration using a Bayesian approach, *J. Hydrol.*, *275*, 12–26, doi:10.1016/S0022-1694(02)00412-2.

Thyer, M., B. Renard, D. Kavetski, G. Kuczera, S. W. Franks, and S. Srikanthan (2009), Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis, *Water Resour. Res.*, *45*, W00B14, doi:10.1029/2008/WR006825.

TNO (2015/2016), DINOlolet. [Available at https://www.dinoloket.nl/ondergrondgegevens.]

Von Asmuth, J. R., and M. F. P. Bierkens (2005), Modeling irregularly spaced residual series as a continuous stochastic process, *Water Resour. Res.*, *41*, W12404, doi:10.1029/2004WR003762.

Von Asmuth, J. R., and M. Knotters (2004), Characterizing groundwater dynamics based on a system identification approach, *J. Hydrol.*, *296*, 118–134, doi:10.1016/j.jhydrol.2004.03.015.

Von Asmuth, J. R., M. F. P. Bierkens, and K. Maas (2002), Transfer function-noise modeling in continuous time using predefined impulse response functions, *Water Resour. Res.*, *38*(12), 1287, doi:10.1029/2001WR001136.

Von Asmuth, J. R., K. Maas, M. Bakker, and J. Petersen (2008), Modeling time series of ground water head fluctuations subjected to multiple stresses, *Ground Water*, *46*, 30–40.

Vrugt, J. A. (2016), Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation, *Environ. Modell. Software*, *75*, 273–316, doi:10.1016/j.envsoft.2015.08.013.

Vrugt, J. A., C. J. F. Ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson (2008), Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, *Water Resour. Res.*, *44*, W00B09, doi:10.1029/2007WR006720.

Yang, J., P. Reichert, K. C. Abbaspour, and H. Yang (2007), Hydrological modelling of the Chaohe Basin in China: Statistical model formulation and Bayesian inference, *J. Hydrol.*, *340*, 167–182, doi:10.1016/j.jhydrol.2007.04.006.

Yapo, P. O., H. V. Gupta, and S. Sorooshian (1996), Automatic calibration of conceptual rainfall-runoff models: Sensitivity to calibration data, *J. Hydrol.*, *181*(1–4), 23–48, doi:10.1016/0022-1694(95)02918-4.