
Tijdreeksanalyse kan veel beter door kleiner modelinterval

Wilbert Berendrecht
Hans Gehrels
Frans van Geer
Arnold Heemink

Bij het analyseren van grondwaterstandsreeksen met behulp van transferfunctie-ruismodellen lijkt het meetinterval soms een beperkende factor te zijn voor een goede beschrijving van het systeem. Door echter gebruik te maken van een toestandsbeschrijving van het systeem hoeft het meetinterval niet meer maatgevend te zijn. Een toestandsbeschrijving gecombineerd met een Kalmanfilter maakt het namelijk mogelijk om het modelinterval los te koppelen van het meetinterval. Dit betekent dus ook dat het modelinterval kleiner kan worden gekozen dan het meetinterval. In dit artikel wordt aangetoond dat dit resulteert in een aanzienlijke verbetering van het model. De mate van verbetering hangt af van het meetinterval en de responskarakteristieken van het te modelleren systeem. Verder laten we zien dat door toevoeging van hoogfrequente waarnemingen het model nog verder verbeterd kan worden.

1 Inleiding

Tijdreeksmodellen zijn een veel gebruikt hulpmiddel bij het analyseren van grondwaterstandsreeksen. Hierbij wordt hoofdzakelijk gebruik gemaakt van transferfunctie-ruis (TFR) modellen zoals beschreven in Box en Jenkins (1970). Een meer generieke manier voor het beschrijven van een systeem is de toestandsbeschrijving. Deze uit de systeemtheorie afkomstige benadering is door Bierkens e.a. (1999) gebruikt voor het analyseren van onregelmatig waargenomen tijdreeksen, iets wat voor de standaard Box-Jenkins-modellen niet mogelijk is. De reden hiervoor is dat in de toestandsbeschrijving het modelinterval niet per definitie gelijk hoeft te zijn aan het meetinterval. Dit betekent dat grondwaterreeksen met een in Nederland gebruikelijke meetfrequentie van 2 waarnemingen per maand, bijvoorbeeld op dagbasis gemodelleerd kunnen worden, natuurlijk wel op voorwaarde dat er dagelijkse gegevens van de invoervariabelen (bijvoorbeeld neerslagoverschot) beschikbaar zijn.

Naast de mogelijkheid om onregelmatige reeksen te analyseren, biedt de ont koppeling van het modelinterval en het meetinterval een aanzienlijk belangrijker voordeel, namelijk de mogelijkheid om grondwatersystemen nauwkeuriger te beschrijven. Een fijnere discreti-

Wilbert Berendrecht en **Arnold Heemink** zijn werkzaam bij TU Delft, afdeling Toegepaste Wiskundige Analyse, Postbus 5031, 2600 GA Delft, e-mail: w.berendrecht@nitg.tno.nl, telefoon: (030) 2564796, fax: (030) 2564755; **Hans Gehrels** is werkzaam bij TU Delft, afdeling Watermanagement, Postbus 5048, 2600 GA Delft; **Frans van Geer** is werkzaam bij TNO-NITG, afdeling Grondwater, Postbus 80015, 3508 TA Utrecht.

satie van de transferfunctie resulteert namelijk in een betere benadering van het onderliggende proces. Concreet betekent dit, dat *zonder hoogfrequente grondwaterstandsmetingen* (zie bijvoorbeeld Van Herwijnen, 2002) het tijdreeksmodel al een betere beschrijving van het systeem geeft.

Om inzicht te krijgen in de daadwerkelijke invloed van een reductie van het modelinterval bij gelijkblijvend meetinterval, hebben we een groot aantal synthetische reeksen geanalyseerd. Door variatie van het modelinterval, het meetinterval en de lengte van de reeksen blijkt in hoeverre en voor welke gevallen een verbetering van het model optreedt. Daarnaast gaan we in op de toepassing van automatische dataloggers. Deze loggers zijn in te stellen op elk gewenst meetinterval en zijn daarom uitermate geschikt om bestaande reeksen met een relatief groot meetinterval uit te breiden met hoogfrequente waarnemingen. Dit artikel beschrijft de resultaten van deze studie. Allereerst wordt kort op de theoretische achtergrond van het gebruikte model ingegaan.

2 Modelbeschrijving

2.1 Toestandsbeschrijving

De algemene toestandsbeschrijving van een SISO (single-input single-output) systeem bestaat uit twee vergelijkingen (Schweppe, 1973):

$$\mathbf{x}_t = \mathbf{A}_t \mathbf{x}_{t-1} + \mathbf{b}_t u_t + \mathbf{g}_t w_t, \quad t = 1, \dots, T, \quad (1)$$

$$y_t = \mathbf{c}_t \mathbf{x}_t + y_r + v_t, \quad t = 1, \dots, T. \quad (2)$$

Hierin is \mathbf{x}_t de toestand van het systeem op tijdstip t ; \mathbf{A}_t is een $m \times m$ transitie matrix; \mathbf{b}_t is een $m \times 1$ vector; u_t is de systeeminvoer op tijdstip t ; \mathbf{c}_t is een $1 \times m$ vector; y_t is de waarneming op tijdstip t ; y_r is een constant, afhankelijk van het referentieniveau van de tijdreeks; en w_t en v_t stellen respectievelijk de systeemruis en de meetruis voor. De beide ruistermen zijn normaal verdeeld met een gemiddelde nul en een varantie

$$\begin{aligned} \text{var}(w_t) &= q, \\ \text{var}(v_t) &= r. \end{aligned} \quad (3)$$

Bovendien zijn beide ruistermen onderling ongecorrleerd. Er zijn diverse mogelijkheden om een TFR-model in bovenstaande vorm te schrijven. Een veel gebruikte vorm is de volgende:

$$\begin{bmatrix} \mathbf{x}_d \\ \mathbf{x}_s \end{bmatrix}_t = \begin{bmatrix} \mathbf{A}_d & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_s \end{bmatrix} \begin{bmatrix} \mathbf{x}_d \\ \mathbf{x}_s \end{bmatrix}_{t-1} + \begin{bmatrix} \mathbf{b}_d \\ \mathbf{0} \end{bmatrix} u_t + \begin{bmatrix} \mathbf{0} \\ \mathbf{g}_s \end{bmatrix} w_t, \quad (4)$$

$$y_t = [\mathbf{c}_d \quad \mathbf{c}_s] \begin{bmatrix} \mathbf{x}_d \\ \mathbf{x}_s \end{bmatrix}_t + y_r + v_t, \quad (5)$$

met

$$\mathbf{A}_d = \begin{bmatrix} \delta_1 & 1 & 0 & \dots & 0 \\ \delta_2 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \delta_{r-1} & \vdots & & \ddots & 1 \\ \delta_r & 0 & \dots & \dots & 0 \end{bmatrix}, \quad \mathbf{b}_d = \begin{bmatrix} \omega_0 \\ \omega_1 \\ \vdots \\ \omega_{r-2} \\ \omega_{r-1} \end{bmatrix}, \quad \mathbf{c}_d = [1 \ 0 \ \dots \ \dots \ 0]$$

$$\mathbf{A}_s = \begin{bmatrix} \phi_1 & 1 & 0 & \dots & 0 \\ \phi_2 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \phi_{p-1} & \vdots & & \ddots & 1 \\ \phi_p & 0 & \dots & \dots & 0 \end{bmatrix}, \quad \mathbf{g}_s = \begin{bmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_{p-2} \\ \theta_{p-1} \end{bmatrix}, \quad \mathbf{c}_s = [1 \ 0 \ \dots \ \dots \ 0].$$

De elementen van \mathbf{x}_d beschrijven de *deterministische* component van het systeem: $x_{d,1} = \eta_t$, $x_{d,2} = \eta_{t,1}$, ..., $x_{d,r} = \eta_{t,r+1}$ terwijl de elementen van \mathbf{x}_s op gelijke wijze de *stochastische* component van het systeem beschrijven: $x_{s,1} = \xi_t$, $x_{s,2} = \xi_{t,1}$, ..., $x_{s,p} = \xi_{t,p+1}$. In een appendix is een voorbeeld opgenomen dat laat zien dat deze toestandsbeschrijving gelijk is aan het TFR-model volgens Box en Jenkins.

De waarneming y_t is dus, naast het referentieniveau y_r en de meetruis v_t , opgebouwd uit η_t en ξ_t . In termen van een grondwatermodel is u het neerslagoverschot, η_t de component van de grondwaterfluctuaties ten gevolge van het neerslagoverschot, ξ_t het residu of onverklaarde deel van de grondwaterfluctuaties en y_r het drainageniveau.

De elementen in bovenstaande matrices zijn onbekend en dienen samen met de variantie q en het referentieniveau y_r geschat te worden. Deze set met te schatten parameters wordt vanaf nu aangeduid met de vector $\alpha = (\delta_1, \dots, \delta_r, \omega_0, \dots, \omega_{r-1}, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_{p-1}, q)$. De variantie van de meetruis r wordt bekend verondersteld.

2.2 Toestands- en parameterschatting

De toestand \mathbf{x}_t kan worden geschat met behulp van het Kalmanfilter. Het Kalmanfilter is een recursief algoritme voor een optimale schatting van de toestandsvector op tijdstip t , op basis van de informatie die op dat tijdstip beschikbaar is. Globaal bestaat het filter uit twee stappen: een voorspelling en een meetcorrectie. Dit is een essentiële eigenschap in het kader van reductie van het modelinterval. Als er namelijk geen meting beschikbaar is wordt de meetcorrectie overgeslagen. Op deze manier kunnen gaten in een meetreeks overbrugd worden.

Het Kalmanfilter wordt geïnitieerd met beginvoorwaarden voor de meetcorrectie, respectievelijk de covariantiematrix van de meetcorrectie:

$$\hat{\mathbf{x}}_0 \text{ en } \mathbf{P}_0. \tag{6}$$

De voorspelling $\bar{\mathbf{x}}_t$ en bijbehorende covariantiematrix \mathbf{M}_t worden als volgt berekend:

$$\begin{aligned} \bar{\mathbf{x}}_t &= \mathbf{A}_t \hat{\mathbf{x}}_{t-1} + \mathbf{b}_t u_t, \\ \mathbf{M}_t &= \mathbf{A}_t \mathbf{P}_t \mathbf{A}_t^T + \mathbf{g}_t q \mathbf{g}_t^T. \end{aligned} \tag{7}$$

Indien een meting beschikbaar komt, kan de schatting van de toestand en bijbehorende covariantiematrix worden gecorrigeerd volgens

$$\begin{aligned}\hat{\mathbf{x}}_t &= \bar{\mathbf{x}}_t + \mathbf{M}_t \mathbf{c}_t^T f_t^{-1} n_t, \\ \mathbf{P}_t &= (\mathbf{I} - \mathbf{M}_t \mathbf{c}_t^T f_t^{-1} \mathbf{c}_t) \mathbf{M}_t.\end{aligned}\quad (8)$$

Hierin is n_t de innovatie op tijdstip t :

$$n_t = y_t - y_r - \mathbf{c}_t \bar{\mathbf{x}}_t, \quad (9)$$

en f_t de innovatievariantie op tijdstip t :

$$f_t = \mathbf{c}_t \mathbf{M}_t \mathbf{c}_t^T + r. \quad (10)$$

Als er geen meting beschikbaar is, wordt de meetcorrectie gelijkgesteld aan de voorspelling:

$$\begin{aligned}\hat{\mathbf{x}}_t &= \bar{\mathbf{x}}_t, \\ \mathbf{P}_t &= \mathbf{M}_t.\end{aligned}\quad (11)$$

De parametervector α kan nu worden geschat met behulp van een 'gereduceerde' log-likelihoodfunctie. Deze functie is gebaseerd op de volgens vergelijking (9) en (10) berekende innovaties en innovatie varianties. De 'gereduceerde' log-likelihoodfunctie wijkt af van de 'gewone' log-likelihoodfunctie doordat zowel $\text{var}(w_t)$ als $\text{var}(v_t)$ worden geschaald volgens

$$\begin{aligned}q_* &= \frac{\text{var}(w_t)}{\sigma_*^2}, \\ r_* &= \frac{\text{var}(v_t)}{\sigma_*^2}.\end{aligned}\quad (12)$$

Stel nu $\sigma_*^2 = \text{var}(w_t)$, dan volgt na enkele stappen de vergelijking voor de 'gereduceerde' log-likelihoodfunctie (Harvey, 1990):

$$\log L(N; \alpha) = -\frac{N}{2} (\log 2\pi + 1) - \frac{1}{2} \sum_{t=1}^N \log f_t(\alpha) - \frac{N}{2} \sum_{t=1}^N \log \sigma_*^2(\alpha), \quad (13)$$

met

$$\sigma_*^2(\alpha) = \frac{1}{N} \sum_{t=1}^N \frac{n_t^2(\alpha)}{f_t(\alpha)}. \quad (14)$$

Hierin is N het aantal waarnemingen. Op deze wijze volgt $q = \text{var}(w_t) = \sigma_*^2(\alpha)$ direct uit het Kalmanfilter. Deze parameter hoeft dus niet meer geschat te worden. De overige parameters worden geschat door maximalisatie van vergelijking (13) of minimalisatie van $-2 \log L(N; \alpha)$. De covariantiematrix van de geschatte parameters wordt benaderd met behulp van de 'Cramer-Rao lower bound' (Schweppe, 1973).

3 Onderzoeksbeschrijving

Om de invloed van een reductie van het modelinterval vast te kunnen stellen is gebruik gemaakt van synthetische reeksen. Het voordeel van het gebruik van synthetische reeksen in plaats van werkelijke reeksen is dat de eigenschappen van deze reeksen exact bekend zijn. In deze paragraaf wordt beschreven hoe de reeksen zijn gegenereerd, welke aspecten zijn onderzocht en op welke wijze de diverse resultaten met elkaar zijn vergeleken.

3.1 Genereren van grondwatertijdreeksen

Evenals in paragraaf 2.1 nemen we aan dat een grondwaterreeks beschreven kan worden door twee componenten: de deterministische component (verklaard uit de invoergegevens) en een stochastische component (onverklaarde deel). Voor het genereren van de tijdreeksen is gebruik gemaakt van dagcijfers van neerslag en Makkink-verdamping, waargenomen in De Bilt door het KNMI. Het hieruit berekende neerslagoverschot P_c (neerslag minus verdamping) is vervolgens getransformeerd met behulp van de kansdichtheidsfunctie van een lognormale verdeling:

$$\Psi_r = \frac{c}{\tau\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln \tau - \mu}{\sigma}\right)^2\right], \quad \sigma > 0, \quad 0 < \tau < \infty, \quad (15)$$

waarin μ en σ respectievelijk het geometrisch gemiddelde en de geometrische standaardafwijking zijn en c een schalingsfactor. De deterministische component van de gegenereerde tijdreeks wordt dan

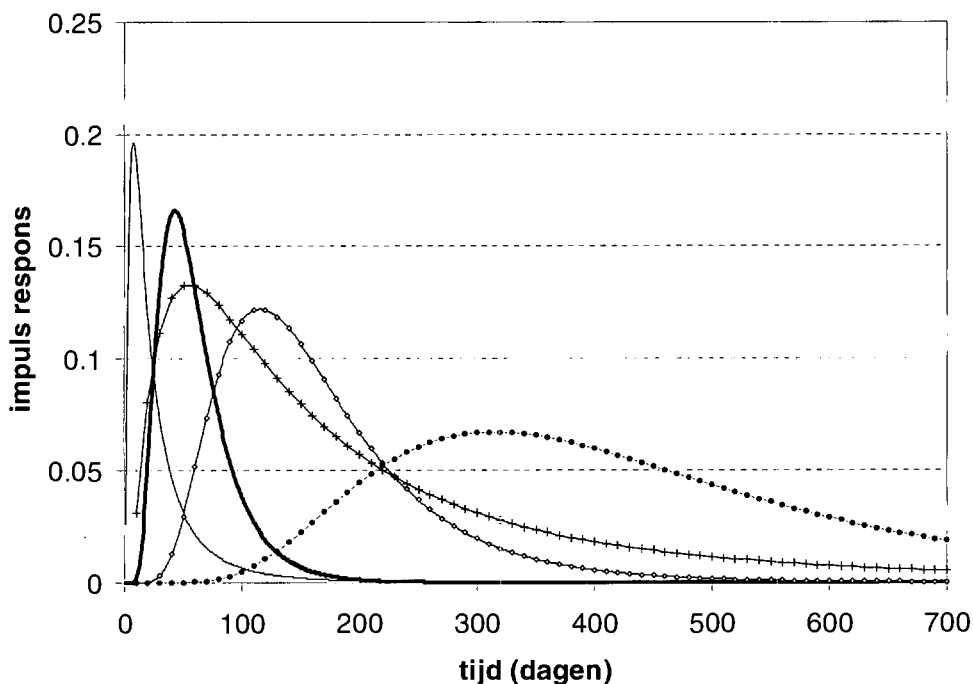
$$\eta_t = \sum_r \Psi_r P_{e,t-r}. \quad (16)$$

De stochastische component wordt beschreven met het volgende autoregressieve model:

$$\xi_t = \phi \xi_{t-1} + w_t, \quad (17)$$

waarin w_t normaal verdeeld is met een gemiddelde van nul en een variantie die gelijk is gesteld aan 1% van de variantie van de deterministische component. Om een statistisch correct experiment uit te voeren, zijn 20 realisaties van vergelijking (17) gebruikt.

Op basis van vergelijkingen (16) en (17) is een groot aantal reeksen gegenereerd voor de periode van 1 juli 1957 tot en met 31 december 1999. De reeksen varieerden van snel (piekrespons binnen een paar dagen) tot langzaam reagerend (piekrespons na een jaar). Enkele responsfuncties die zijn gebruikt voor het genereren van deze reeksen zijn weergegeven in figuur 1. Aangezien de conclusies voor de verschillende typen reeksen niet fundamenteel verschillen, zullen alleen de resultaten van de tijdreeks gegenereerd met de vetgedrukte responscurve (piekrespons op tijdstip $t_p = 43$ dagen) worden besproken. Deze deterministi



Figuur 1: Enkele responsfuncties die gebruikt zijn voor het genereren van tijdreeksen. In dit artikel wordt de vetgedrukte curve als voorbeeld gebruikt.

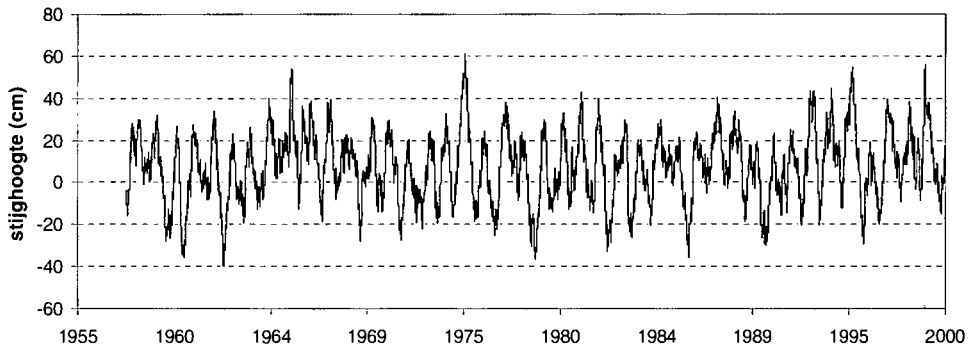
Tabel 1: Gebruikte parameters voor het genereren van de tijdreeksen.

Reeks	Deterministische component			Stochastische component	
	μ	σ	c	ϕ	S_N
R1	4,0	0,5	10	0	0,01
R2	4,0	0,5	10	0,99	0,37

sche component is gecombineerd met twee verschillende stochastische componenten. In tabel 1 zijn de gebruikte parameterwaarden weergegeven. Hierin is S_N gedefinieerd als

$$S_N = \frac{\text{var}(\xi_t)}{\text{var}(\eta_t)} \quad (18)$$

Figuur 2 toont een realisatie van tijdreeks R2. Om het gekalibreerde model te kunnen valideren is de tijdreeks opgeknipt in een kalibratieperiode (1957–1989) en een validatieperiode (1990–1999).



Figuur 2: Realisatie van een gegenereerde tijdreeks. Voor de deterministische component is gebruik gemaakt van een lognormale kansdichtheidsfunctie, de stochastische component is gegenereerd op basis van een autoregressief model (zie reeks R2 in tabel 1).

Tenslotte zijn diverse ‘samples’ van de reeksen gegenereerd. Hierbij werden de volgende aspecten gevarieerd:

1 *Meetinterval*

Het meetinterval is gevarieerd tussen 10 en 70 dagen: $dt_{meet} = 10, 20, \dots, 70$ dagen.

2 *Lengte van de reeks*

Het is te verwachten dat de invloed van intervalreductie afhangt van de lengte van de tijdreeks. In dit experiment zijn dan ook naast de lange reeks van 33 jaar, korte tijdreeksen geanalyseerd: lengte $l = 10, 20, \dots, 100\%$ van de maximale lengte (33 jaar).

3 *Aanvulling meetreeks met hoogfrequente waarnemingen*

In dit onderzoek zijn reeksen geanalyseerd waarbij het eerste stuk van de reeks bestaat uit waarnemingen met een interval van 70 dagen en het tweede stuk uit waarnemingen met een interval van 10 dagen. Ook hierbij is de lengte van de reeks gevarieerd.

3.2 *Evaluatiecriteria*

Om de invloed van een reductie van het modelinterval vast te stellen zijn evaluatiecriteria nodig. Als eerste criterium is gekozen voor de ‘fit’ van de deterministische component. Dankzij het gebruik van synthetische reeksen is deze fit namelijk exact te bepalen. De fit geeft aan hoe goed de deterministische component kan worden gescheiden van de stochastische component. Een parameter die deze fit (of beter gezegd: afwijking) goed beschrijft, is de *mean absolute error* MAE [L]:

$$MAE = \frac{1}{N} \sum_{t=1}^N |\hat{\eta}_t - \eta_t|, \quad (19)$$

met $\hat{\eta}_t$ de geschatte deterministische component. De MAE kwantificeert dus hoe goed de deterministische component gescheiden kan worden van de stochastische component.

Het tweede criterium is de nauwkeurigheid waarmee de transferfunctie geschat kan worden. Deze nauwkeurigheid wordt goed uitgedrukt door de standaardafwijking van de zogenaamde *gain*: s_G [$L(LT^{-1})^{-1}$]. Hierbij is de *gain* G [$L(LT^{-1})^{-1}$] gedefinieerd als het oppervlak onder de transferfunctie:

$$G = \frac{\sum_i \omega_i}{1 - \sum_k \delta_k} \quad (20)$$

4 Invloed van gereduceerd modelinterval

In deze paragraaf wordt vastgesteld wat het effect is van een gereduceerd modelinterval op de uitkomsten van het tijdreeksmodel. Daartoe worden steeds twee typen modellen met elkaar vergeleken:

- Model 1: het modelinterval is gelijk aan het meetinterval;
- Model 2: het modelinterval is 10 dagen.

Hierbij wordt gebruik gemaakt van de twee tijdreeksen uit tabel 1: reeks R1 met een witte ruiscomponent (geen autoregressieve term) en reeks R2 met een autoregressieve ruiscomponent.

4.1 Variatie van meetinterval

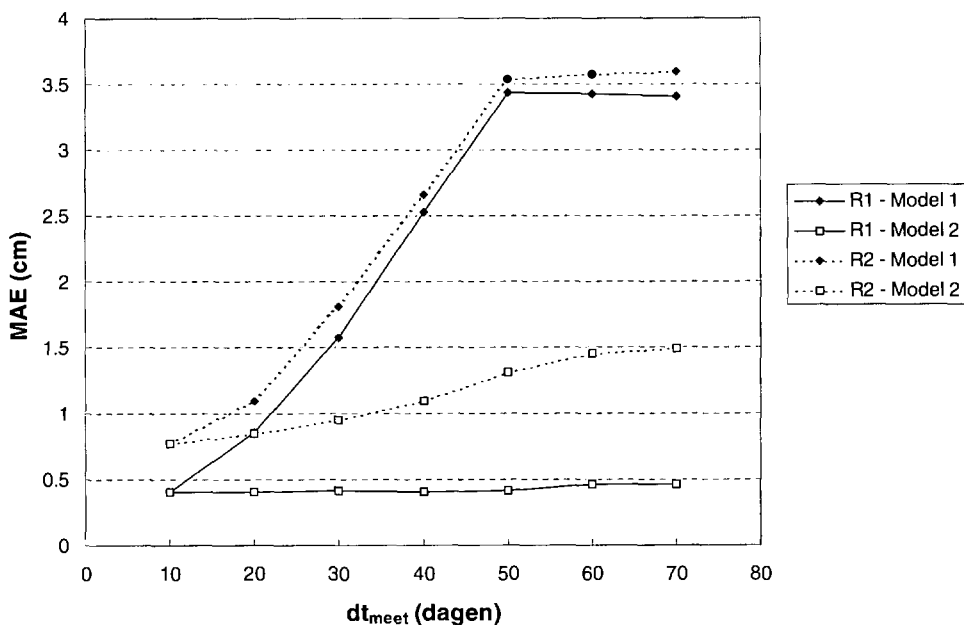
In figuur 3 is de fit van de deterministische component uitgezet tegen het meetinterval. De doorgetrokken curven tonen de MAE voor reeks R1, de onderbroken curven de MAE voor reeks R2. Bij gebruik van model 1 vertonen beide reeksen een zelfde verloop: als het meetinterval toeneemt, neemt de fout ook toe. Pas nadat het meetinterval groter is dan het tijdstip van piekrespons $t_p = 43$ dagen, neemt de fout nauwelijks meer toe. Daarnaast ligt de curve van R2 iets hoger dan die van R1. Dit komt doordat de stochastische component van R2 aanzienlijk groter is. Dit bemoeilijkt de schatting van de transferfunctie.

Veel interessanter is natuurlijk het verschil tussen model 2 en model 1. De figuur laat namelijk zien dat model 2 een veel betere fit geeft dan model 1. Voor reeks R1 geldt zelfs dat het meetinterval er nauwelijks meer toe doet: zelfs bij een meetinterval van 70 dagen is de MAE maar 0,5 cm. Dit is een reductie ten opzichte van model 1 van 85%. Hoewel voor reeks R2 de MAE nog wel toeneemt met het meetinterval, is ook hier de verbetering ten opzichte van model 1 aanzienlijk.

Bovenstaande kan worden verklaard door beschouwing van de transferfunctie. Het modelinterval bepaalt namelijk de discretisatie van deze functie: hoe kleiner het modelinterval, des te fijner de discretisatie en dus ook des te beter de benadering van de transferfunctie.

Het tweede evaluatiecriterium is de standaardafwijking van de gain s_c . In figuur 4 is dit criterium uitgezet tegen het meetinterval. Voor de volledigheid: de gain van de transferfunctie is gelijk aan c in tabel 1: 10 cm per 1 mm/dag. Dit betekent dus dat bij een continu neerslagoverschot van 1 mm/dag de grondwaterstand uiteindelijk 10 cm stijgt.

Figuur 4 laat zien dat s_c op gelijke wijze afhangt van het meetinterval als MAE. Bij model 1 stijgt de standaardafwijking bij een toename van het meetinterval. Ook nu vlakkt de curve af (zij het in mindere mate) als het meetinterval groter wordt dan t_p . Een belangrijk verschil met MAE is dat s_c veel gevoeliger is voor de stochastische component: de standaardafwijking voor reeks R2 is aanzienlijk hoger dan die voor reeks R1. Hierdoor



Figuur 3: Relatie tussen de fit van de deterministische component (*mean absolute error*, MAE) en het meetinterval voor reeks R1 (ongecorreleerde stochastische component) en R2 (autoregressieve stochastische component). Voor model 1 is het modelinterval gelijk aan het meetinterval, terwijl model 2 een vast modelinterval van 10 dagen heeft.

is bij reeks R2 de relatieve invloed van een gereduceerd modelinterval kleiner dan bij reeks R1.

Stelling 1:

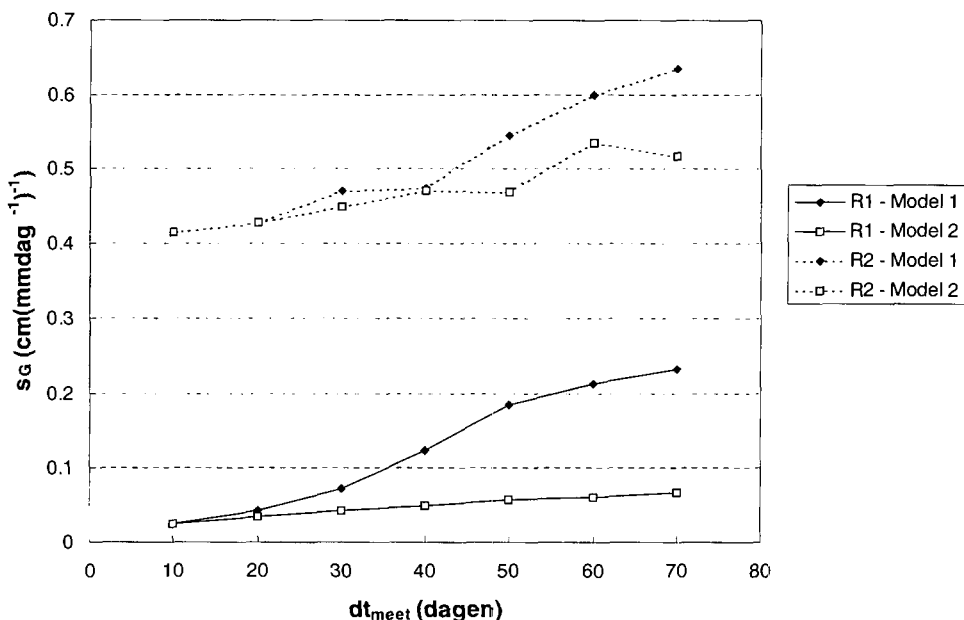
Een reductie van het modelinterval resulteert, ongeacht het meetinterval, in een betere fit van de deterministische component en een nauwkeurigere schatting van de transferfunctie.

Stelling 2:

Bij gebruik van een gereduceerd modelinterval neemt de gevoeligheid voor het meetinterval af. Hierbij speelt de stochastische component een belangrijke rol: een grotere stochastische component resulteert in een grotere gevoeligheid.

Stelling 3:

De relatieve invloed van een gereduceerd modelinterval op de nauwkeurigheid van de geschatte transferfunctie wordt sterk bepaald door de stochastische component.

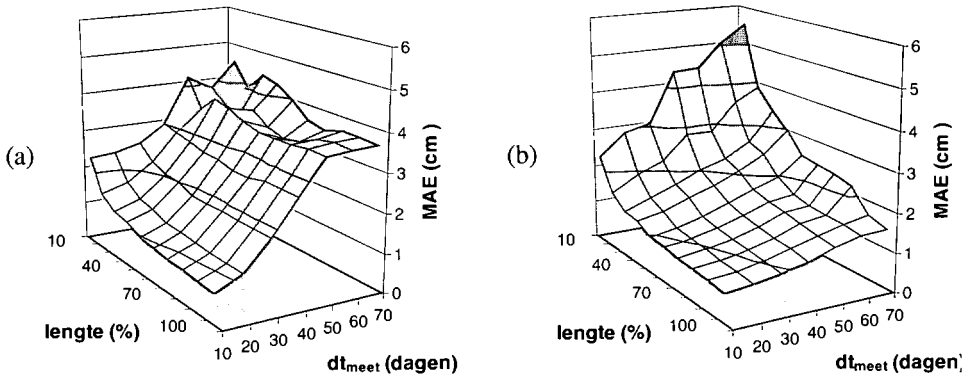


Figuur 4: Relatie tussen de nauwkeurigheid van de transferfunctie (standaardafwijking van de gain s_G) en het meetinterval voor reeks R1 (ongecorreleerde stochastische component) en R2 (autoregressieve stochastische component). Voor model 1 is het modelinterval gelijk aan het meetinterval, terwijl model 2 een vast modelinterval van 10 dagen heeft.

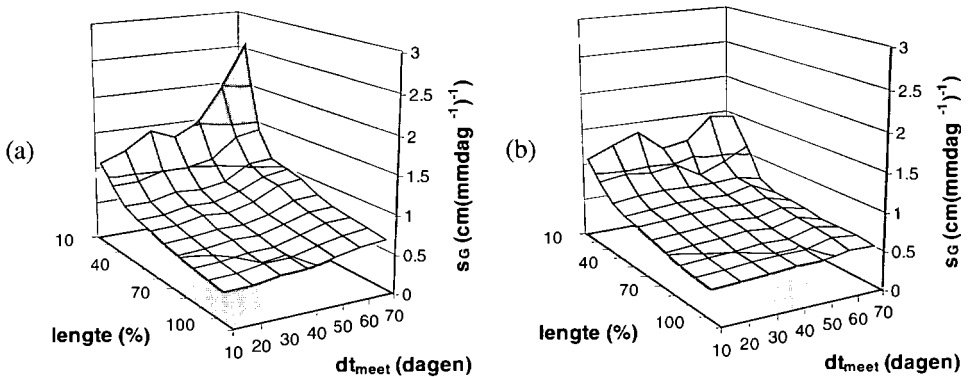
4.2 Variatie van reekslengte

Om het overzicht te bewaren wordt de invloed van de reekslengte alleen gepresenteerd voor reeks R2. Figuur 5 toont de fit van de deterministische component als een functie van zowel de relatieve reekslengte (ten opzichte van de oorspronkelijke lengte van 33 jaar) als het meetinterval. Voor een lengte van 100% is de curve dus gelijk aan die in figuur 3. Uit figuur 5 blijkt dat de reekslengte pas echt invloed krijgt op MAE als de lengte kleiner wordt dan 50% (16,5 jaar) van de oorspronkelijke lengte. Deze lengte hangt uiteraard nauw samen met de responstijd. Bij sneller reagerende systemen zal deze lengte kleiner zijn. Opmerkelijk is dat de MAE van model 2 (figuur 5b) bij korte reeksen en grote meetintervallen nauwelijks meer afwijkt van de MAE van model 1. Dit is eenvoudig te verklaren uit het feit dat in deze gevallen zeer weinig waarnemingen beschikbaar zijn. Door gebruik van een klein modelinterval is het aantal modelparameters groter en is de kans op 'overfitting' groot. Het wordt dan dus moeilijk om de deterministische component te scheiden van de stochastische component.

In tegenstelling tot MAE hangt de nauwkeurigheid van de geschatte transferfunctie (s_G) in sterke mate af van de reekslengte (zie figuur 6). Dit heeft te maken met de relatie tussen de parameter nauwkeurigheid en het aantal waarnemingen. De parameter nauwkeurigheid wordt nog weleens globaal evenredig gesteld met $n^{-0.5}$, met n het aantal waarnemingen. Deze relatie is aardig goed terug te vinden in figuur 6a. Tenslotte toont figuur 6 dat de reekslengte het effect van een reductie van het modelinterval nauwelijks beïnvloedt.



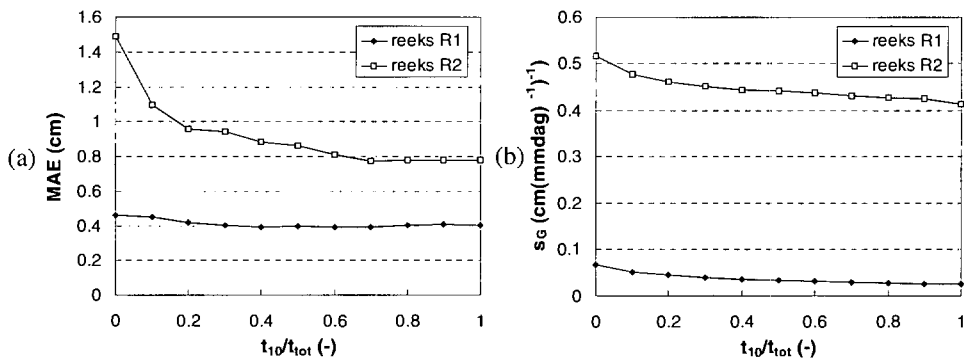
Figuur 5: Fit van de deterministische component (MAE) in de relatie tot het meetinterval en de reekslengte (percentage van de oorspronkelijke lengte van 33 jaar) voor reeks R2 en (a) model 1 met een modelinterval gelijk aan het meetinterval; en (b) model 2 met een modelinterval van 10 dagen.



Figuur 6: Nauwkeurigheid van de geschatte transferfunctie (s_c) in relatie tot het meetinterval en de reekslengte (percentage van de oorspronkelijke lengte van 33 jaar) voor reeks R2 en (a) model 1 met een modelinterval gelijk aan het meetinterval; en (b) model 2 met een modelinterval van 10 dagen.

4.3 Toevoegen van hoogfrequente waarnemingen

Voorgaande resultaten hebben laten zien dat een reductie van het modelinterval al een aanzienlijk betere beschrijving van het systeem oplevert. De vraag is nu of hoogfrequente waarnemingen hier nog iets aan toevoegen. In figuur 7 is daarom weergegeven wat de invloed is van het toevoegen van hoogfrequente waarnemingen op de fit van de deterministische component (MAE) en de nauwkeurigheid van de transferfunctie (s_c). Op de horizontale as is de fractie hoogfrequente waarnemingen uitgezet. Bij $t_{10}/t_{tot} = 1$ bestaat de gehele reeks uit 10-daagse waarnemingen, terwijl bij $t_{10}/t_{tot} = 0$ de gehele reeks uit 70-daagse waarnemingen bestaat. Het modelinterval is gelijk aan het laagste meetinterval: 10 dagen. Als eerste valt op dat (vooral in figuur 7a) de invloed van hoogfrequente waarnemingen



Figuur 7: Invloed van toegevoegde hoogfrequente waarnemingen (laatste t_{10}/t_{tot} fractie van de reekslengte) aan reeks R1 en R2 op (a) de fit van de deterministische component (MAE); en (b) de nauwkeurigheid van de geschatte transferfunctie (s_G). Het modelinterval is 10 dagen en de reekslengte is 33 jaar.

het grootste effect heeft bij reeks R2. Dit is ook logisch, want het verschil tussen MAE bij $t_{10}/t_{tot} = 0$ en $t_{10}/t_{tot} = 1$ is veel groter. Ten tweede is duidelijk te zien dat de eerste hoogfrequente waarnemingen (10-20%) het grootste effect hebben op MAE. Dit is vanuit praktisch oogpunt erg aantrekkelijk. De curve van s_G laat een meer gelijkmatige daling zien. Dit komt omdat s_G meer afhangt van het aantal dan van het meetinterval.

Stelling 4:

De overstap van laagfrequent naar hoogfrequent waarnemen resulteert al snel in een aanzienlijke verbetering van de fit van de deterministische component.

5 Test case

In de vorige paragraaf is vastgesteld dat tijdreeksmodellen aanzienlijk verbeterd kunnen worden door reductie van het modelinterval. Tevens is er vastgesteld dat een reductie van het meetinterval verder kan bijdragen aan de verbetering van het model. In deze paragraaf zullen de resultaten worden getoetst aan een bestaande grondwaterreeks.

5.1 Beschrijving dataset

De gebruikte grondwaterreeks (code 33gp023401) is waargenomen in een freatisch filter in de buurt van Eerbeek. De meetfrequentie bedraagt 24 waarnemingen per jaar en de reeks loopt van 1990 tot 2000. Deze reeks is uitermate geschikt als test case vanwege het feit dat in 1999 de meetfrequentie voor een periode van drie maanden (dit is ongeveer 2,5% van de totale reekslengte) is verhoogd naar 1 waarneming per dag. Om het gekalibreerde model te kunnen valideren is de reeks opgesplitst in een kalibratieperiode (1990 tot juni 1999) en een validatieperiode (juli 1999 tot oktober 2000). De dagcijfers voor neerslag komen van het KNMI-station Eerbeek, terwijl de dagcijfers voor verdamping van het KNMI-hoofdstation in De Bilt komen.

Voor deze *test case* werden twee samples geselecteerd:

- 1 meetinterval van 14 dagen, exclusief dagwaarnemingen,

2 meetinterval van 14 dagen, inclusief dagwaarnemingen.

Bovendien werd het eerste 'sample' met twee verschillende modellen geanalyseerd:

1a modelinterval van 14 dagen (model 1 uit de vorige paragraaf),

1b modelinterval van 1 dag (model 2 uit de vorige paragraaf).

Uiteraard is het modelinterval voor 'sample' 2 ook 1 dag.

5.2 Modelresultaten

Voor alle drie modellen bleek de volgende modelvorm het meest geschikt:

$$\mathbf{x}_t = \begin{bmatrix} \delta_1 & 1 & 0 \\ \delta_2 & 0 & 0 \\ 0 & 0 & \phi_1 \end{bmatrix} \mathbf{x}_{t-1} + \begin{bmatrix} \omega_1 \\ \omega_2 \\ 0 \end{bmatrix} u_t + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} w_t, \quad (21)$$

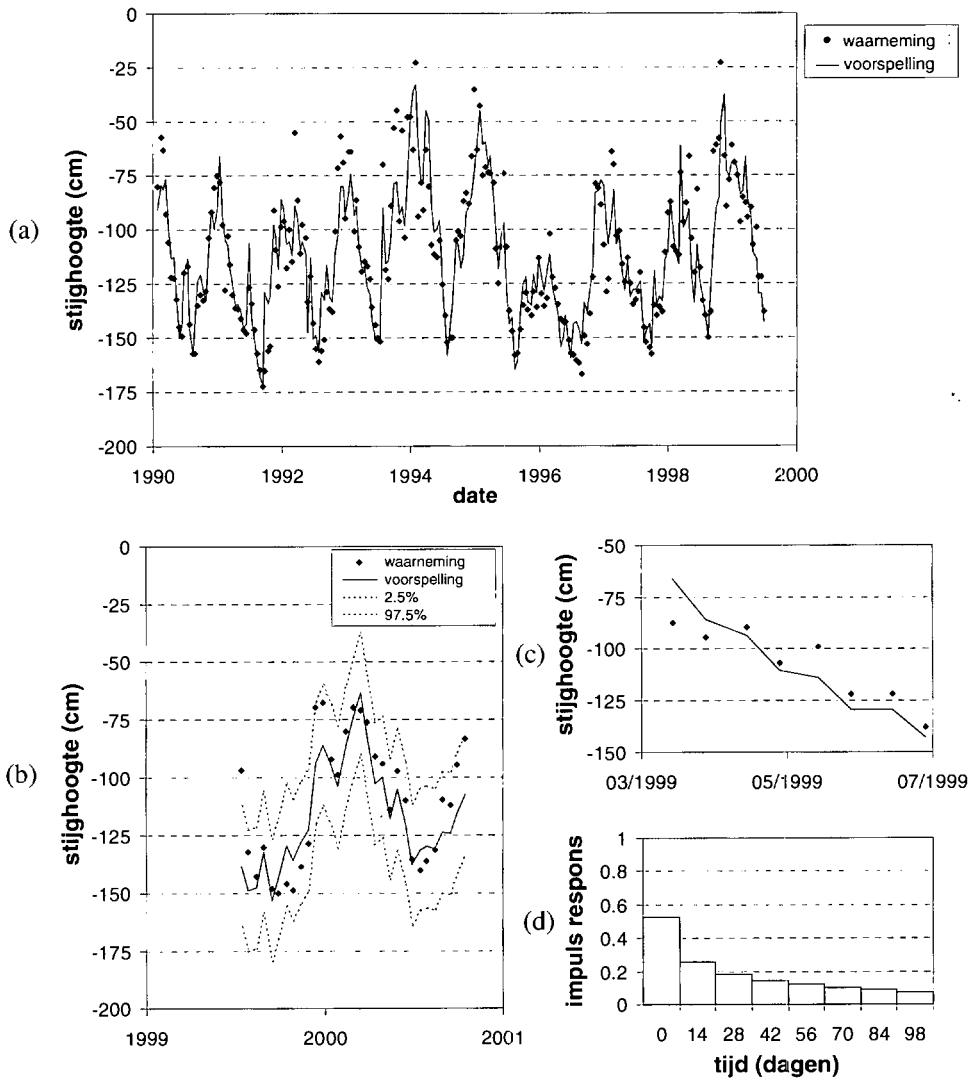
$$y_t = [1 \quad 0 \quad 1] \mathbf{x}_t + y_r + v_t.$$

In tabel 2 zijn de geschatte parameters van de drie modellen weergegeven. De resultaten zijn weergegeven in figuur 8 (model 1a) en 9 (model 2). Aangezien de resultaten van model 1b en 2 erg op elkaar lijken (zie tabel 2), zijn alleen de resultaten van model 2 afgebeeld.

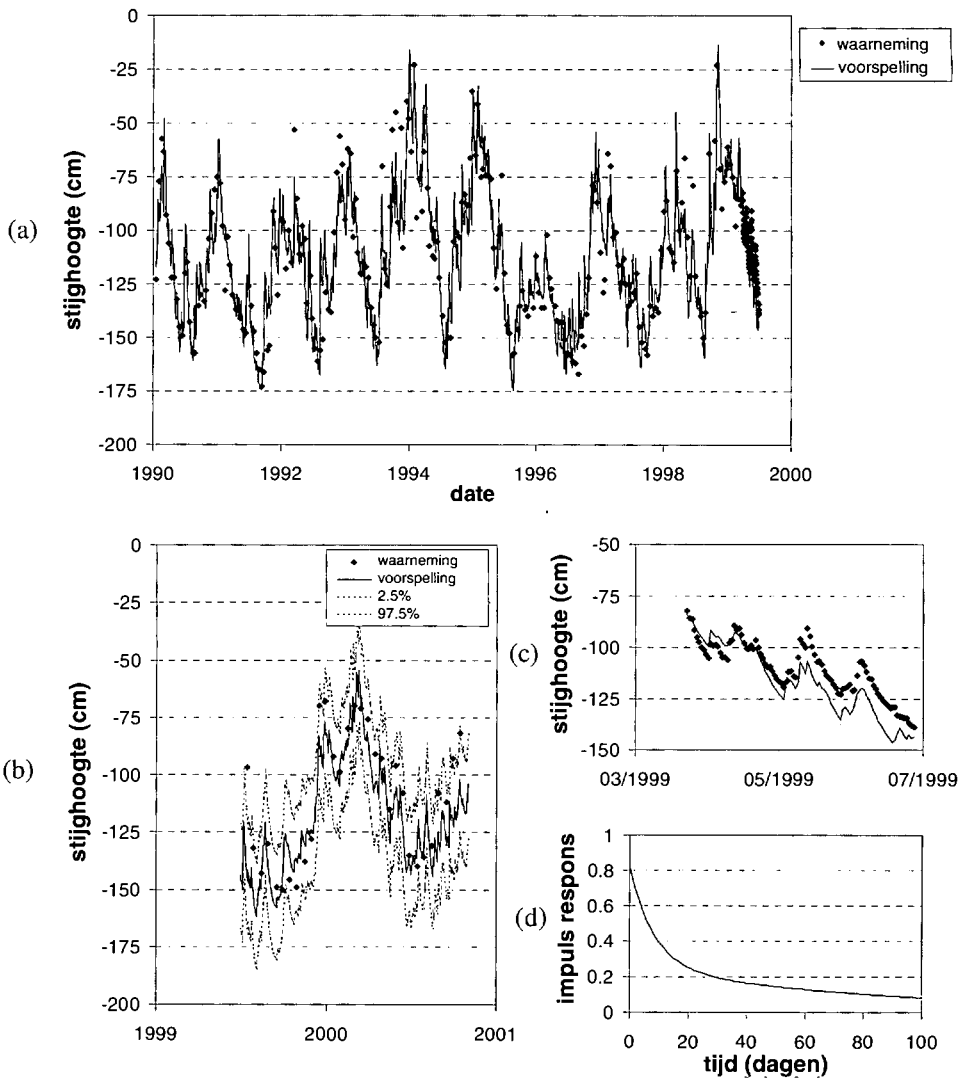
Tabel 2: Geschatte parameters van de gekalibreerde modellen. De standaardafwijking van de schattingfout is tussen haakjes weergegeven.

model	dt_{meer}	dt_{mod}	δ_1	δ_2	ϕ_1	ω_0	ω_1	y_r
	<i>dag</i>	<i>dag</i>	-	-	-	<i>cm</i>	<i>cm</i>	<i>cm</i>
						<i>(mmdag⁻¹)'</i>	<i>(mmdag⁻¹)'</i>	
1a	14	14	1,0438 (0,12)	-0,1698 (0,092)	0,3906 (0,044)	7,377 (0,020)	-4,066 (0,056)	-138,6 (1,9)
1b	14	1	1,8594 (0,015)	-0,8609 (0,015)	0,9417 (0,0068)	0,8724 (0,035)	-0,8317 (0,034)	-140,1 (1,6)
2	gemengd	1	1,8729 (0,011)	-0,8742 (0,010)	0,9578 (0,0044)	0,8264 (0,028)	-0,7905 (0,027)	-139,8 (1,8)

Op basis van beide figuren blijkt allereerst dat een klein modelinterval het mogelijk maakt om de piekrespons veel beter te benaderen. Hierdoor zijn de pieken in de reeks veel beter te modelleren. Tevens blijkt uit figuur 9c dat model 2 de dagelijks waargenomen fluctuaties behoorlijk goed beschrijft. Dit in tegenstelling tot model 1a (modelinterval van 14 dagen), dat eigenlijk alleen maar het voortschrijdende gemiddelde beschrijft. Dit is ook logisch, omdat het neerslagoverschot over een periode van 14 dagen wordt uitgesmeerd.



Figuur 8: Resultaten van model 1a; (a) waarnemingen en voorspellingen voor de kalibratieperiode; (b) waarnemingen en voorspellingen met 95% betrouwbaarheidsinterval voor de validatieperiode; (c) waarnemingen en voorspellingen voor de periode met dagelijkse waarnemingen (ter vergelijking met figuur 9); en (d) geschatte impuls-responsfunctie.



Figuur 9: Resultaten van model 2: (a) waarnemingen en voorspellingen voor de kalibratieperiode; (b) waarnemingen en voorspellingen met 95% betrouwbaarheidsinterval voor de validatieperiode; (c) waarnemingen en voorspellingen voor de periode met dagelijkse waarnemingen; en (d) geschatte impuls-responsfunctie.

De drie modellen kunnen verder worden vergeleken op basis van enkele criteria. De standaardafwijking van de gain (s_e) kan weer gebruikt worden, maar de fit van de deterministische component kan nu niet meer berekend worden. Daarom wordt gebruik gemaakt van enkele andere criteria: de variantie van de innovaties van het gekalibreerde ($n_{c,t}$) en het gevalideerde ($n_{v,t}$) deel en de variantie van de stochastische component of 'residu' (ξ_t). Deze criteria zijn opgenomen in tabel 3.

Tabel 3: Vergelijking van de drie modellen op basis van enkele evaluatiecriteria

model	df_{meer}	df_{mod}	$var(n_{c,v})$	$var(n_{v,v})$	$var(\xi)$	G	s_c
	dag	dag	cm ²	cm ²	cm ²	cm (mmdag ⁻¹) ⁻¹	cm (mmdag ⁻¹) ⁻¹
1a	14	14	152.7	118.2	180.5	26.29	1.8
1b	14	1	120.7	116.7	142.0	26.74	1.4
2	gemengd	1	89.2	114.8	136.7	26.50	1.4

Op basis van de tabel kunnen twee belangrijke conclusies worden getrokken. Allereerst blijkt dat de reductie van het modelinterval resulteert in reductie van $var(\xi)$ van ongeveer 21%. Indien de dagelijkse waarnemingen worden meegenomen (model 2) daalt de variantie nog eens verder met 3%. Ten tweede daalt de standaardafwijking van de gain met ruim 22%, wat overeenkomt met de bevindingen in de vorige paragraaf. De verlaging van s_c door toevoeging van dagelijkse waarnemingen is echter nog niet merkbaar. Gezien het feit dat de set dagelijkse waarnemingen slechts 2,5% van de totale reekslengte uitmaakt, was dit ook te verwachten.

Samenvattend kan dus worden gesteld dat deze test case de bevindingen in de vorige paragrafen bevestigen: door reductie van het modelinterval geeft het model een betere beschrijving van de grondwaterreeks.

6 Enkele belangrijke conclusies

De belangrijkste conclusie van dit onderzoek is dat tijdreeksmodellen aanzienlijk verbeterd kunnen worden door toepassing van een kleiner modelinterval. Dit kan door gebruik te maken van de generieke toestandsbeschrijving van een systeem. De verbetering van het model komt tot uitdrukking in een betere procesbeschrijving (een betere benadering van de transferfunctie) en een nauwkeuriger schatting van de transferfunctie. Een reductie van het modelinterval is het meest effectief als het meetinterval groot is ten opzichte van de responstijd van het systeem. Een verdere verbetering van het model kan worden bereikt door toepassing van hoogfrequente (ten opzichte van de responstijd) waarnemingen. Verder tonen de resultaten van dit onderzoek aan dat bij trage reeksen de huidige meetfrequentie van 2 waarnemingen per maand zonder problemen verlaagd kan worden. Zolang het modelinterval onveranderd blijft, zal deze verlaging nauwelijks invloed hebben op de modelresultaten.

Een ander groot voordeel van de toestandsbeschrijving is dat het model eenvoudig kan worden uitgebreid naar meer complexe modellen. Zo kunnen extra niet waargenomen toestanden worden toegevoegd. Ook niet-lineaire relaties kunnen worden ingebouwd. Inmiddels richt het onderzoek zich dan ook op het mee-modelleren van de wortelzone. Het op deze wijze ontstane niet-lineaire model blijkt (vooral bij diepe reeksen) goede resultaten te geven. Hier zal in een vervolgartikel verder op worden ingegaan.

Referenties

- Bierkens, M.F.P., M. Knotters en F.C. van Geer (1999)** Tijdreeksanalyse nu ook toepasbaar bij onregelmatige meetfrequenties; in: *Stromingen*, jrg 5, nr 2, pag. 43–54.
- Box, G.E.P. en G.M. Jenkins (1970)** Time series analysis, forecasting and control; Holden-Day, San Fransisco.
- Harvey, A.C. (1990)** Forecasting, structural time series models and the Kalman filter; Cambridge University Press, Cambridge.
- Herwijnen, T. van (2002)** Tijdreeksanalyse van grondwaterpeilen: de invloed van het meetinterval; in: *Stromingen*, jrg 8, nr 4, pag. 19–30.
- Scheppe, F.C. (1973)** Uncertain dynamic systems; Prentice-Hall, New Jersey.

Appendix: herschrijven van toestandsbeschrijving in Box–Jenkins-vorm

Beschouw als voorbeeld de volgende toestandsbeschrijving, bestaande uit een systeemvergelijking en een meetvergelijking:

$$\begin{bmatrix} x_{d,1} \\ x_{d,2} \\ x_{d,3} \\ x_{s,1} \\ x_{s,2} \end{bmatrix}_t = \begin{bmatrix} \delta_1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \phi_1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_{d,1} \\ x_{d,2} \\ x_{d,3} \\ x_{s,1} \\ x_{s,2} \end{bmatrix}_{t-1} + \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \\ 0 \\ 0 \end{bmatrix} u_t + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ \theta_1 \end{bmatrix} w_t, \quad (\text{A1})$$

$$y_t = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_{d,1} \\ x_{d,2} \\ x_{d,3} \\ x_{s,1} \\ x_{s,2} \end{bmatrix}_t + y_r + v_t.$$

De systeemvergelijking kan worden geschreven als het volgende stelsel vergelijkingen:

$$\begin{aligned} x_{d,1}(t) &= \delta_1 x_{d,1}(t-1) + x_{d,2}(t-1) + \omega_1 u(t), \\ x_{d,2}(t) &= x_{d,3}(t-1) + \omega_2 u(t), \\ x_{d,3}(t) &= \omega_3 u(t), \\ x_{s,1}(t) &= \phi_1 x_{s,1}(t-1) + x_{s,2}(t-1) + w(t), \\ x_{s,2}(t) &= \theta_1 w(t). \end{aligned} \quad (\text{A2})$$

Volledig uitgeschreven staat hier:

$$\begin{aligned} x_{d,1}(t) &= \delta_1 x_{d,1}(t-1) + \omega_3 u(t-2) + \omega_2 u(t-1) + \omega_1 u(t), \\ x_{s,1}(t) &= \phi_1 x_{s,1}(t-1) + \theta_1 w(t-1) + w(t). \end{aligned} \quad (\text{A3})$$

Gecombineerd met de meetvergelijking

$$y(t) = x_{d,1}(t) + x_{s,1}(t) + y_r + v_t \quad (\text{A4})$$

is te zien dat vergelijking (A1) een TFR(1,3,1,1)-model is. Het enige verschil met een Box-Jenkins TFR-model is dat vergelijking (A1) nog een term voor de meetruis bevat.