Validatie van tijdreeksmodellen voor de grondwaterstand

Martin Knotters[1]

'Essentially, all models are wrong, but some are useful', is een gezegde van de statisticus George E.P. Box, die samen met Gwilym M. Jenkins het bekende boek 'Time Series Analysis: Forecasting and Control' schreef. In dit artikel wil ik ingaan op de pragmatische vraag of een tijdreeksmodel bruikbaar is voor een bepaald doel. Ik beperk me tot tijdreeksmodellen voor grondwaterstandsreeksen. Na een korte inleiding over tijdreeksmodellering van grondwaterstandsreeksen zal ik ingaan op de verschillende manieren waarop de kwaliteit van deze modellen wordt uitgedrukt, wat dit zegt over de bruikbaarheid van een model, wanneer je zou kunnen spreken van 'validatie' en hoe je met validatie een idee kunt krijgen van de bruikbaarheid van een tijdreeksmodel. Vervolgens zal ik ter illustratie validatieresultaten presenteren voor tijdreeksmodellen die zijn gekalibreerd op vijftien reeksen van grondwaterstanden, om af te sluiten met enkele concluderende opmerkingen.

Inleiding

Tijdreeksmodellen worden al decennia lang gefit op waarnemingsreeksen van Nederlandse grondwaterstanden, voor uiteenlopende doeleinden. Paul Baggelaar (1988) en Frans van Geer (1988) pasten vierentwintig jaar geleden bijvoorbeeld al tijdreeksmodellering toe om het effect van grondwaterwinning op de grondwaterstand te kwantificeren. Harry Rolf (1989) paste tijdreeksmodellering toe om trends in grondwaterstandswaarnemingen in de periode 1950-1986 te analyseren. Hans Gehrels (1995) paste tijdreeksmodellering toe bij het analyseren van verlagingen van de grondwaterstanden op de Veluwe.
Tijdreeksmodellen zijn regressiemodellen, die de relatie beschrijven tussen één of meer verklarende variabelen, de input, en een responsvariabele, de output. Naar het handboek van Box en Jenkins (1970) worden tijdreeksmodellen ook wel Box-Jenkinsmodellen genoemd. In een 'univariaat' tijdreeksmodel wordt de waarde van een variabele op een bepaald tijdstip verklaard uit waarden van die variabele op voorgaande tijdstippen. We spreken dan van 'auto-regressie', regressie-met-zichzelf. Bij tijdreeksmodellen voor de grondwaterstand maken we echter meestal ook gebruik van één of meer 'exogene' variabelen, zoals neerslag- en verdampingscijfers.

[1] Alterra, onderdeel van Wageningen UR, martin.knotters@wur.nl

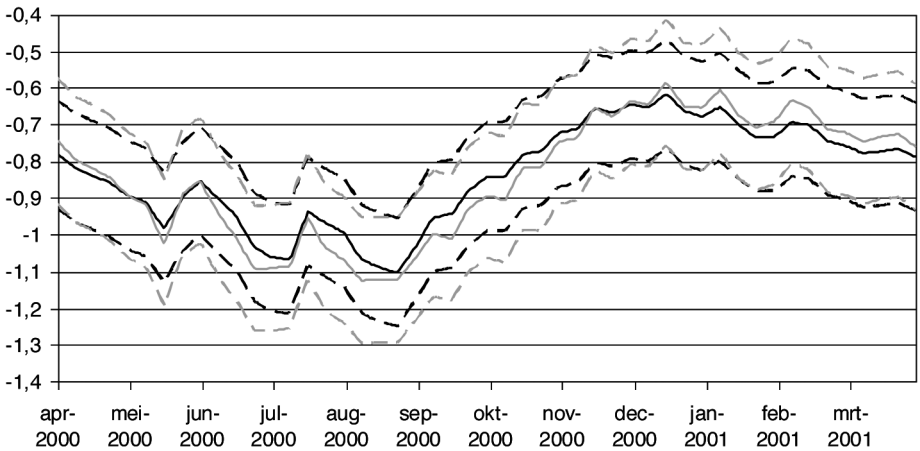
Tijdreeksmodellen konden aanvankelijk alleen worden gekalibreerd op reeksen met gelijke waarnemingsintervallen (equidistante reeksen). Equidistante reeksen van grondwaterstanden kwamen in de praktijk vaak alleen bij benadering voor: halfmaandelijks waargenomen grondwaterstanden zijn niet precies equidistant, zoals automatische waarnemingen op dag- of uurfrequentie dat wel kunnen zijn. Door het Kalmanfilteralgoritme toe te passen (Bierkens e.a., 1999; Berendrecht e.a., 2003) en door te werken met continue tijd (Von Asmuth e.a., 2001; 2002) kan tijdreeksmodellering ook op niet-equidistante reeksen worden toegepast.
Tijdreeksmodellen worden ook wel 'black-boxmodellen' genoemd, omdat zij empirische relaties beschrijven tussen waarnemingsreeksen, en niet expliciet zijn gebaseerd fysische wetmatigheden zoals fysisch-mechanistische modellen dat wel zijn, en die daarom ook wel 'white-boxmodellen' worden genoemd. Tussen black- en white-boxmodellen liggen de grey-boxmodellen: modellen met een statistische basis en een meer of minder gedetailleerde fysische onderbouwing. Voorbeelden van grey-boxmodellen voor de grondwaterstand zijn het model dat is gebaseerd op een stochastische differentiaalvergelijking (Bierkens, 1998), het gecombineerde bodem-grondwatermodel met stochastische invoer (EMERALD; Bierkens en Walvoort, 1998), het fysisch-gebaseerde ARX(1,0)-model (Knotters en Bierkens, 1999) en het PIRFICT-model dat deel uitmaakt van het pakket Menyanthes (Von Asmuth e.a., 2001, 2002; Von Asmuth, 2012).
Met tijdreeksmodellen is het mogelijk de nauwkeurigheid van voorspellingen te kwantificeren. Deze gekwantificeerde nauwkeurigheid kan bijvoorbeeld worden gebruikt om de kans op overschrijding van een kritisch niveau te schatten, of om te laten zien wat de kwaliteit is van een voorspelling. Informatie over de nauwkeurigheid kan op diverse manieren worden gepresenteerd. Er kan bijvoorbeeld een 95%-voorspellingsinterval worden aangegeven rond een grondwaterstandsverloop dat met een tijdreeksmodel is gesimuleerd (Afbeelding 1).
De gekwantificeerde voorspelfout en de 95%-voorspellingsintervallen die ervan zijn afgeleid zijn echter schattingen. Het is informatie over de nauwkeurigheid, gegeven het model. Onzekerheid over het model is er niet in verdisconteerd, zie McLeod en Hipel (1978) en Chatfield (1995). Zou je voor een ander model hebben gekozen, dan zou je andere resultaten krijgen met andere 95%-voorspellingsintervallen, zoals Afbeelding 1 laat zien. Of een tijdreeksmodel geschikt is voor bijvoorbeeld het berekenen van overschrijdingskansen hangt niet alleen af van de nauwkeurigheid van de voorspellingen, maar ook van de kwaliteit van de gekwantificeerde nauwkeurigheid.
Doel van dit artikel is om te bespreken hoe je de bruikbaarheid van een tijdreeksmodel kunt beoordelen en welke rol de vergelijking van modeluitkomsten met additionele, onafhankelijke waarnemingen daarbij kan spelen. Ik beperk me in dit artikel tot de kwaliteit van tijdreeksmodellen die zijn gekalibreerd op grondwaterstandswaarnemingen. De kwaliteit van de waarnemingen zelf komt in dit artikel niet aan de orde.

Kwaliteit van tijdreeksmodellen

Goodness-of-fit

Als je een tijdreeksmodel hebt gekalibreerd op een reeks, dan zou je een maat die de 'goodness-of-fit' uitdrukt kunnen gebruiken als kwaliteitslabel. Dit kan een absolute maat zijn, zoals het gemiddelde van de gekwadrateerde afwijkingen tussen

de gefitte en de waargenomen grondwaterstanden, of de wortel uit dit gemiddelde (= root mean squared error, RMSE). Of het kan een relatieve maat zijn, zoals het percentage verklaarde variantie. Dit geeft aan hoeveel procent van de variatie van de grondwaterstand wordt verklaard door het model. Stel je voor dat het model niets anders is dan een constant niveau, ook al variëren de waarnemingen nog zo: in dat geval is het percentage verklaarde variantie 0%. Het andere uiterste is dat het model perfect samenvalt met de waarnemingen: in dat geval is het percentage verklaarde variantie 100%.



Afbeelding 1: Tijdreeksgrafiek van gesimuleerde grondwaterstanden (lijnen) en 95%-voorspellingsintervallen (streeplijnen) in m t.o.v. NAP voor de locatie B11C0329_1. Zwart: simulatie met model PIRFICT, niet-lineair. Grijs: simulatie met model PIRFICT, lineair. De modellen zijn gekalibreerd op gegevens uit de periode 2002-2010.

Diagnostic checks

De 'goodness-of-fit' geeft aan hoe goed het model bij de waarnemingen past. Met 'diagnostic checks' kun je controleren of het model voldoet aan de onderliggende veronderstellingen. Je kunt bijvoorbeeld controleren of de residuen die na het kalibreren overblijven ongecorreleerd zijn, en een constante variatie hebben. Is er nog correlatie in de residuen aanwezig, dan beschrijft het model blijkbaar nog niet alle samenhang en kan het model nog worden verbeterd. Ook mogen de residuen niet gecorreleerd zijn met de gefitte waarden, want dit kan bij simulaties problemen opleveren. Handboeken over tijdreeksmodellering, zoals Box en Jenkins (1970) en Hipel en McLeod (1994) beschrijven uitvoerig deze diagnostic checks.

Bruikbaarheid

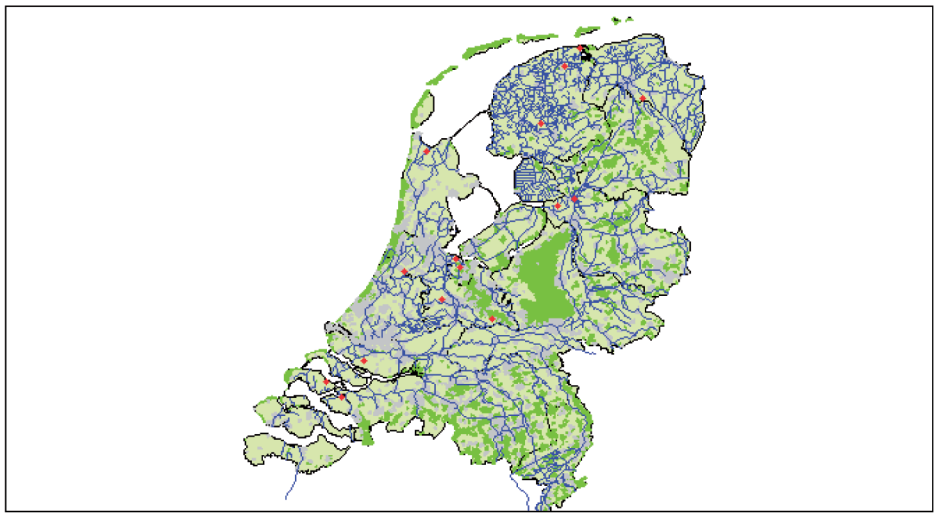
De 'goodness-of-fit' en 'diagnostic checks', zeggen iets over hoe goed het model bij de waarnemingen past en in hoeverre het model beantwoordt aan de onderliggende veronderstellingen. Ze geven echter geen antwoord op de vraag hoe bruikbaar het model is voor een bepaald doel. Hoe erg is het dat het model niet perfect fit en dat de ruis niet helemaal wit is? Om daar achter te komen zou je de proef op de som moeten nemen.

Dat kan door modeluitkomsten te vergelijken met onafhankelijke waarnemingen, dat wil zeggen waarnemingen die niet zijn gebruikt bij de modelselectie en de modelkalibratie. Een dergelijke proef wordt vaak validatie genoemd. De term 'validatie' wordt in de literatuur echter verschillend gehanteerd. Chatfield (1995) gebruikt de term voor 'model checking' in het algemeen, maar benadrukt daarbij het belang van een aanvullende set onafhankelijke waarnemingen, om ook modelonzekerheid goed te kunnen kwantificeren. Oreskes e.a. (1994) gebruiken de termen 'verification', 'validation' en 'confirmation', en stellen dat verificatie en validatie van numerieke modellen onmogelijk is. Verificatie houdt in hun artikel in dat de 'truth', waarheid, van een model wordt aangetoond, terwijl validatie betekent dat de geldigheid van een model wordt vastgesteld. De discussie over de onmogelijkheid van verificatie en validatie blijkt voor een groot deel te gaan over de pretenties die aan resultaten van verificatie en validatie worden verbonden. Oreskes e.a. (1994) merken op dat veel van wat doorgaat voor 'verificatie' of 'validatie' hooguit 'bevestiging' is: in hoeverre worden modeluitkomsten bevestigd door waarnemingen? Belangrijk is om zich daarbij te realiseren dat je beperkt bent tot wat waarneembaar is, dat de waarnemingen een steekproef uit een populatie vormen en dat de waarnemingen zelf een bepaalde nauwkeurigheid hebben.
Ik sluit in dit artikel aan bij de definitie die Bohlin (1991, blz. 76) van validatie geeft: een testprocedure, waarvan de uitkomst aangeeft of een model voldoet aan zijn doel. Dit is een pragmatische aanpak, waarbij modellering niet tot doel heeft alle data te verklaren (de 'scientist's rule'), maar modellering erop is gericht een bepaald doel te dienen (de 'engineer's rule'). Het validatiecriterium op basis waarvan de bruikbaarheid van het model wordt beoordeeld moet dus verband houden met het doel van het model. Bij doelen kun je bijvoorbeeld denken aan het voorspellen van toekomstige grondwaterstanden, het simuleren van reeksen om bijvoorbeeld overschrijdingskansen te schatten, of om als invoer voor scenarioberekeningen te dienen, het kwantificeren van invloeden, etc.

Illustratie: beoordeling van tijdreeksmodellen voor vijftien reeksen

Ter illustratie heb ik op vijftien tijdreeksen van grondwaterstandswaarnemingen een aantal tijdreeksmodellen gekalibreerd met het pakket Menyanthes (Von Asmuth, 2012), en vervolgens een aantal validaties uitgevoerd. De reeksen zijn gebruikt door Knotters e.a. (2011) en ontleend aan het DINO-loket van TNO. Afbeelding 2 geeft de ligging van de buizen waarin de reeksen zijn waargenomen weer, en tabel 1 geeft informatie over de gegevens waarop de modellen zijn gekalibreerd. De reeksen vallen alle in de periode januari 2002 - januari 2012, al varieert de lengte onderling wat. Ook de meetfrequenties variëren: bij sommige reeksen is op zeker moment de frequentie verhoogd naar dagfrequentie na het plaatsen van een druksensor. Op twee manieren zijn onafhankelijke validatiesets verkregen: 1) door reeksen doormidden te splitsen; 2) door een klein aantal waarnemingen aselect uit de reeks te trekken en deze niet te gebruiken tijdens de kalibratie. De reeksen zijn gesplitst op 16 mei 2006, dat is gemiddeld genomen het middelpunt van de vijftien reeksen. De validatiesets die door aselecte trekking zijn gevormd bestaan uit 24 waarnemingen, waarbij er twaalf zijn getrokken uit het zomerhalfjaar en 12 uit het winterhalfjaar. Het aantal van 24 is arbitrair, misschien wat aan de krappe kant, maar zelfs bij de reeksen met de minste waarnemingen bleven nog voldoende waarnemingen over om met enige nauwkeurigheid een model te kalibreren. De enige reden om de 24 waarnemingen aselect te trekken is objectiviteit, zodat me

later niet kan worden aangewreven dat ik de resultaten die mij het beste uitkomen bij elkaar heb gezocht.



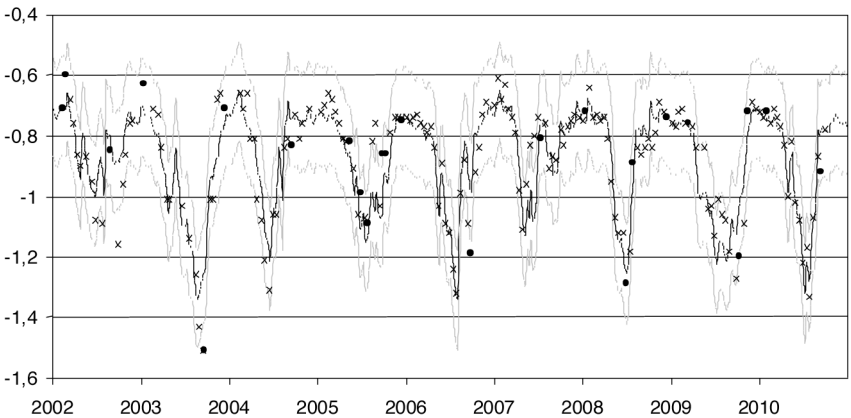
Afbeelding 2: Ligging van de buizen waarin de vijftien grondwaterstandsreeksen zijn waargenomen.

Filternr.	Neerslagstation, verdampingsstation	Niet-lineair	Aantal waarnemingen in de kalibratiesets			
			2002- 2011	2002-2011, excl. 24 wn.	2002-medio 2006	medio 2006-2011
B02G0367_1	Ezumazijl, De Kooij	Ja	168	144	83	85
B06B0139_1	Dokkum, De Kooij	Nee	3195	3171	1597	1598
B11C0329_1	Leeuwarden, De Bilt	Ja	187	163	85	102
B12E0446_1	Onnen, De Bilt	Nee	178	154	91	87
B14B0162_1	Anna-Paulowna, De Kooij	Nee	185	161	99	86
B21D0566_1	IJsselmuiden, De Bilt	Nee	2274	2250	625	1649
B21E0171_1	Rouveen, De Bilt	Nee	194	170	100	94
B25H0659_1	Weesp, De Bilt	Nee	206	182	101	105
B25H0734_2	Weesp, De Bilt	Nee	206	182	101	105
B31A0103_1	Hoofddorp, De Bilt	Nee	899	875	103	796
B31G0304_1	Vleuten, De Bilt	Nee	184	160	82	102
B39B0476_1	De Bilt, De Bilt	Nee	192	168	89	103
B43C0369_1	Noordgouwe, Vlissingen	Ja	197	173	91	106
B43D0295_1	Anna-Jacobapolder, Vlissingen	Nee	195	171	90	105
B43E0281_2	Numansdorp, De Bilt	Nee	164	140	87	77

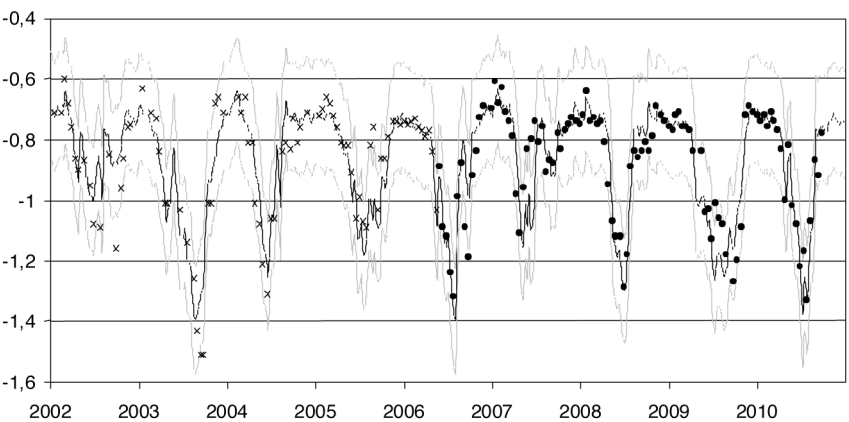
Tabel 1: Informatie over de reeksen waarop PIRFICT-modellen zijn gekalibreerd. Ook is aangegeven of een drempel-nietlineariteit in het model is opgenomen.

Samengevat zijn er zijn modellen gekalibreerd op:

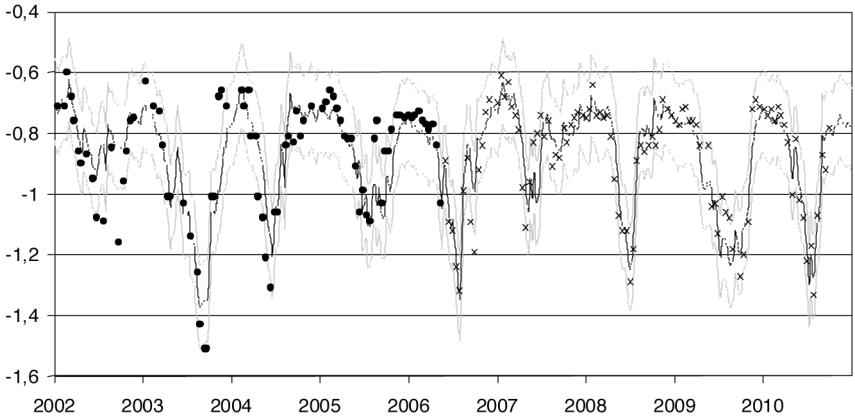
1) alle data in de periode 1 januari 2002 t/m 31 januari 2011. Dan blijft er dus niets over om te valideren;
2) alle data exclusief 24 aselect getrokken data, waarvan 12 in het zomerhalfjaar vallen en 12 in het winterhalfjaar. De 24 data die apart zijn gezet kunnen worden gebruikt voor validatie. Afbeelding 3 geeft hiervan een voorbeeld voor reeks B11C0329_1;
3) de data in de periode 1 januari 2002 t/m 16 mei 2006. De data uit de periode 17 mei 2006 t/m 31 januari 2011 kunnen worden gebruikt voor validatie. Afbeelding 4 geeft hiervan een voorbeeld;
4) de data in de periode 17 mei 2006 t/m 31 januari 2011. De data uit de periode 1 januari 2002 t/m 16 mei 2006 kunnen worden gebruikt voor validatie. Afbeelding 5 geeft hiervan een voorbeeld.



Afbeelding 3: Validatieset verkregen door 24 data aselect te trekken (stippen). Het model is gekalibreerd op de resterende 163 data (kruisjes). Zwarte lijn: gesimuleerd grondwaterstandsverloop. Grijze lijnen: boven- en ondergrens van het 95%-voorspellingsinterval. Grondwaterstanden in m t.o.v. NAP uitgezet tegen de tijd. Locatie: B11C0329_1.



Afbeelding 4: Validatieset verkregen door splitsing van de reeks. Het model is gekalibreerd op 85 waarnemingen (kruisjes) en gevalideerd op 102 waarnemingen (stippen). Zwarte lijn: gesimuleerd grondwaterstandsverloop. Grijze lijnen: boven- en ondergrens van het 95%-voorspellingsinterval. Grondwaterstanden in m t.o.v. NAP uitgezet tegen de tijd. Locatie: B11C0329_1.



Afbeelding 5: Validatieset verkregen door splitsing van de reeks. Het model is gekalibreerd op 102 waarnemingen (kruisjes) en gevalideerd op 85 waarnemingen (stippen). Zwarte lijn: gesimuleerd grondwaterstandsverloop. Grijze lijnen: boven- en ondergrens van het 95%-voorspellingsinterval. Grondwaterstanden in m t.o.v. NAP uitgezet tegen de tijd. Locatie: B11C0329_1.

Tabel 2 geeft informatie over de 'goodness-of-fit' van de modellen. Duidelijk is dat een hoog percentage verklaarde variantie niet een kleine RMSE hoeft te betekenen, het zijn immers resp. een relatieve en een absolute maat.

Filternr.	Kalibratieset							
	2002- 2011		2002-2011, excl. 24 wn.		2002-medio 2006		medio 2006-2011	
	% v.v.	RMSE	% v.v.	RMSE	% v.v.	RMSE	% v.v.	RMSE
B02G0367_1	58	10,7	54	10,2	63	8,8	63	11,1
B06B0139_1	92	11,0	91	11,5	89	12,0	94	10,2
B11C0329_1	85	7,3	82	7,7	80	8,7	85	6,8
B12E0446_1	70	12,2	67	12,0	74	11,5	65	12,9
B14B0162_1	69	13,6	67	14,3	68	11,4	77	13,7
B21D0566_1	84	4,1	84	4,1	81	4,0	86	4,2
B21E0171_1	58	5,8	56	6,1	76	3,2	59	6,7
B25H0659_1	81	3,1	80	3,1	79	3,1	79	3,3
B25H0734_2	70	6,9	71	6,7	71	6,5	70	7,1
B31A0103_1	71	21,4	69	21,9	79	12,8	82	19,0
B31G0304_1	77	9,6	76	9,8	83	8,4	73	10,3
B39B0476_1	65	16,5	64	17,2	87	8,8	60	19,1
B43C0369_1	86	14,5	85	14,4	89	13,2	82	15,2
B43D0295_1	73	12,5	74	11,9	76	12,0	70	12,8
B43E0281_2	83	11,9	81	12,1	85	10,2	82	13,6

Tabel 2: Maten voor goodness-of-fit van het model PIRFICT, gekalibreerd op vijftien reeksen van waargenomen grondwaterstanden.% v.v. = percentage verklaarde variantie, RMSE = root mean squared error (cm).

Bij de validatie zijn met de gekalibreerde modellen grondwaterstandsreeksen gesimuleerd en deze zijn vergeleken met de waargenomen grondwaterstanden in de validatiesets. De validatiecriteria moeten verband houden met het doel van het model, teneinde de bruikbaarheid te kunnen benutten. Stel dat het doel vrij algemeen is, bijvoorbeeld om

lange reeksen te simuleren als invoer voor een scenariostudie of om statistieken zoals overschrijdingskansen uit te berekenen. Je zou dan voor de validatieset de verschillen kunnen berekenen tussen de waargenomen grondwaterstand h en de gesimuleerde grondwaterstand $\hat{h}$ op tijdstip $i, i = 1 \ldots n$, met n het aantal waarnemingen in de validatieset:

$$e_i = h_i - \hat{h}_i.$$

Deze verschillen kunnen worden samengevat met een gemiddelde fout als maat voor de systematische fout:

$$ME = \frac{1}{n}\sum_{i=1}^{n} e_i \ ,$$

de standaardafwijking van de fout als maat voor de toevallige fout:

$$SDE = \sqrt{\frac{\sum_{i=1}^{n}(e_i - ME)^2}{n-1}}$$

en met de root mean squared error, die zowel de systematische als de toevallige fout bevat:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}} = \sqrt{ME^2 + \frac{(n-1)}{n} SDE^2} \ .$$

Als je wilt weten hoe goed het 95%-voorspellingsinterval de nauwkeurigheid van de gesimuleerde grondwaterstandsreeks weergeeft, kan je voor de validatieset bepalen hoeveel procent van de waarnemingen er binnen en buiten dit interval vallen. Dit zou resp. rond de 95 en 5% moeten liggen.

Tabel 3 geeft validatieresultaten in termen van ME, RMSE en percentage waarnemingen buiten het 95%-voorspellingsinterval, voor de twee validatiesets die door splitsing van reeksen tot stand zijn gekomen. Opvallend zijn de relatief grote systematische fouten die bij een aantal reeksen optreden. Dit kan betekenen dat er een ingreep in de waterhuishouding is geweest, of dat de kalibratieperiode onvoldoende informatie over de respons van de grondwaterstand op neerslag en verdamping bevatte.
Verder laat Tabel 3 zien dat bij een aantal modellen het percentage waarnemingen buiten het 95%-voorspellingsinterval groter is dan 5%. Dit is te verwachten aangezien in de 95%-voorspellingsintervallen niet de modelonzekerheid is verdisconteerd (zie McLeod en Hipel, 1978, en Chatfield, 1995). Opmerkelijker is dat bij een aantal modellen het percentage waarnemingen buiten het 95%-voorspellingsinterval kleiner is dan 5%. Hier lijkt de onzekerheid te worden overschat. Dit duidt er op dat de variatie van de residuen niet constant is in de tijd.

Reeksnr.	Kalibratieperiode 17-5-2006 t/m 31-12-2010 Validatieperiode 1-1-2002 t/m 16-5-2006			Kalibratieperiode 1-1-2002 t/m 16-5-2006 Validatieperiode 17-5-2006 t/m 31-12-2010		
	ME	RMSE	P	ME	RMSE	P
B02G0367_1	0,7	12,4	11	0,9	13,4	12
B06B0139_1	-3,0	13,0	12	0,4	13,1	5
B11C0329_1	0,2	9,7	16	-0,2	7,7	5
B12E0446_1	-2,8	11,9	3	2,1	13,2	1
B14B0162_1	0,7	15,1	6	3,6	17,1	16
B21D0566_1	-2,9	4,7	5	1,5	4,9	8
B21E0171_1	-5,1	6,8	3	4,8	8,6	34
B25H0659_1	0,5	3,2	5	0,1	3,4	6
B25H0734_2	-1,5	6,9	4	1,7	7,4	10
B31A0103_1	18,7	28,5	19	-21,7	38,8	35
B31G0304_1	1,1	9,9	2	-0,5	11,0	7
B39B0476_1	-11,3	15,5	0	13,2	23,6	26
B43C0369_1	-5,3	16,0	3	7,3	16,8	10
B43D0295_1	-3,7	12,8	4	4,9	13,8	8
B43E0281_2	-3,2	11,2	1	2,5	14,7	22

Tabel 3: Validatieresultaten, voor validatiesets die zijn verkregen door reeksen te splitsen. ME = gemiddelde fout (cm), RMSE = root mean squared error (cm), P = het percentage waarnemingen buiten het 95%-voorspellingsinterval.

Reeksnr.	ME	RMSE	P
B02G0367_1	-2,7	17,7	17
B06B0139_1	0,7	13,2	8
B11C0329_1	-0,9	10,1	13
B12E0446_1	4,7	13,3	4
B14B0162_1	-0,8	8,2	0
B21D0566_1	-1,0	3,4	0
B21E0171_1	-0,1	4,1	0
B25H0659_1	0,4	2,5	0
B25H0734_2	2,3	8,7	13
B31A0103_1	-15,6	33,6	17
B31G0304_1	-1,0	7,5	0
B39B0476_1	-3,7	10,8	0
B43C0369_1	-1,0	14,9	4
B43D0295_1	-5,6	12,8	8
B43E0281_2	-2,8	10,8	4

Tabel 4: Validatieresultaten voor de validatiesets die zijn verkregen door aselecte trekking (n=24, 12 in het zomerhalfjaar, 12 in het winterhalfjaar). ME = gemiddelde fout (cm), RMSE = root mean squared error (cm), P = het percentage waarnemingen buiten het 95%-voorspellingsinterval.

Tabel 4 geeft validatieresultaten voor de validatieset die is verkregen door aselecte trekking van 24 data. Hier valt de grote systematische fout op in de voorspellingen bij reeks B31A0103_1. Verder liggen ook hier bij een aantal reeksen aanmerkelijk meer dan 5% van de waarnemingen buiten het 95%-voorspellingsinterval. Nu is de omvang van de validatieset vrij gering (24), maar grote afwijkingen ten opzichte van 5% mogen misschien toch te denken geven. Opmerkelijk is dat het model voor B31A0103_1 nog

redelijk goed bij de waarnemingen past (percentage verklaarde variantie = 71%, en 69% als 24 waarnemingen zijn weggelaten), maar dat de validatieresultaten grote systematische voorspelfouten laten zien en een mogelijke onderschatting van de breedte van het 95%-voorspellingsinterval.

De validatieresultaten in tabel 3 en 4 zijn uitgedrukt in ME, RMSE en het percentage waarnemingen buiten het 95%-voorspellingsinterval. Deze maten geven nuttige nformatie over de bruikbaarheid van het tijdreeksmodel bij het simuleren van grondwaterstandsreeksen. Voor meer specifieke doelen is het raadzaam om validatiematen te definiëren die in direct verband staan met dit doel. Het aantal mogelijkheden is schier onbeperkt. Om een idee te geven: in Knotters en De Gooijer (1999) vergeleken we gesimuleerde en waargenomen overschrijdingsfrequenties met elkaar.

Enkele conclusies

Uit Tabel 2 t/m 4 blijkt dat een goede fit niet altijd hoeft samen te gaan met een hoge nauwkeurigheid van gesimuleerde grondwaterstanden en van 95%-voorspellingsintervallen. De bruikbaarheid van tijdreeksmodellen voor het simuleren van lange reeksen kan dus niet uitsluitend worden beoordeeld op basis van informatie over de 'goodness-of-fit'.
Systematische voorspelfouten die door validatie aan het licht worden gebracht geven aan dat niet is voldaan aan stationariteitsveronderstellingen, of dat het model de dynamische relatie tussen neerslag, verdamping en grondwaterstand niet goed beschrijft. Dit laatste kan bijvoorbeeld optreden als de kalibratieperiode niet de responstijd omvat, als de waarnemingen in de kalibratieperiode niet het volledige fluctuatietraject vertegenwoordigen (onvoldoende spreiding in de eigenschappenruimte), of als het model niet alle invloeden bevat.
Chatfield (1995) geeft de voorkeur aan validatie op basis van een volledig nieuwe set gegevens, boven het splitsen van reeksen of apart zetten van gegevens (hold-out samples). Als nadelen van de laatste methoden noemt hij dat je moet beslissen over hoe je de reeks splitst, of over hoeveel data je apart zet voor validatie. Verder noemt hij als nadeel het verlies van efficiëntie omdat je niet alle beschikbare data gebruikt voor kalibratie van het model. Tegelijkertijd merkt hij op dat het bij validatie van tijdreeksmodellen vaak onvermijdelijk is om te werken met een 'hold-out sample'. Validatie op basis van een volledig nieuwe set gegevens zou bij tijdreeksmodellen voor grondwaterstanden betekenen dat je een reeks die op een andere locatie is waargenomen gebruikt. De waterhuishoudkundige omstandigheden op die locatie zullen echter zelden exact overeenkomen met die op de locatie waar de kalibratiegegevens zijn verzameld, wat interpretatie van de validatieresultaten bemoeilijkt.
Validatie door een reeks doormidden te splitsen mag dan niet altijd aantrekkelijk zijn omdat liefst alle data worden benut om het model te kalibreren, toch kan het inzicht opleveren in bijvoorbeeld de mate waarin stationariteit van een regime kan worden verondersteld.
Validatie door aselect een deel van de waarnemingen te trekken en apart te zetten tijdens de kalibratie is nuttig als je wilt weten of een model bruikbaar is voor het opvullen van gaten in een reeks. Vooral nu er hoogfrequent wordt gemeten hoeft het uitsluiten van een deel van de waarnemingen niet te leiden tot een veel slechtere fit van

het model. Het is aan te bevelen om met experimenten te analyseren welke omvang van de validatieset nauwkeurige validatieresultaten oplevert zonder belangrijk verlies van goodness-of-fit.
Validatiecriteria moeten informeren over de bruikbaarheid van een model, en daarom moeten deze criteria verband houden met het doel van het model. Daarom is het aan te bevelen om niet alleen naar ME's en RMSE's te kijken, maar ook naar de kwaliteit van bijvoorbeeld geschatte overschrijdingsfrequenties, voorspellingen van grondwaterstanden in het voorjaar, etc.

Literatuur

Asmuth, J.-R. von (2012) Groundwater System Identification through Time Series Analysis; Ph.D. thesis Delft University of Technology.

Asmuth, J. von, K. Maas en M.F.P. Bierkens (2001) Waarom doen alsof de neerslag eens per maand valt? Het discrete Box-Jenkins- versus het continue PIRFICT-tijdreeksmodel, in theorie; Stromingen; 7(4), pag 33-44.

Asmuth, J. von, M.F.P. Bierkens en K. Maas (2002) Soms is weten beter dan meten (tenzij je verkeerd zit natuurlijk). Het discrete Box-Jenkins versus het continue PIRFICT transferruis-model, in praktijk; Stromingen; 8(1), pag 5-14.

Baggelaar, P.K. (1988) Tijdreeksanalyse bij verlagingsonderzoek: principe en voorbeeld; H_2O; 21(16), pag 443-450.

Berendrecht, W., H. Gehrels, F. van Geer en A. Heemink (2003) Tijdreeksanalyse kan veel beter door kleiner modelinterval; Stromingen; 9(1), pag 5-22.

Bierkens, M.F.P. (1998) Eenvoudige stochastische modellen voor grondwaterstandsfluctuaties. Deel 1: Een stochastische differentiaalvergelijking; Stromingen; 4(2), pag 5-26.

Bierkens, M.F.P. en D.J.J. Walvoort (1998) Eenvoudige stochastische modellen voor grondwaterstandsfluctuaties. Deel 2: Gecombineerd bodem-grondwatermodel met stochastische invoer. Stromingen; 4(3), pag 5-20.

Bierkens, M.F.P., M. Knotters en F.C. van Geer (1999) Tijdreeksanalyse nu ook toepasbaar bij onregelmatige meetfrequenties; Stromingen; 5(2), pag 43-54.

Bohlin, T. (1991) Interactive system identification: prospects and pitfalls; Springer, Berlijn.

Box, G.E.P. en G.M. Jenkins (1970) Time Series Analysis: Forecasting and Control; Holden-Day, San Francisco.

Chatfield, C. (1995) Model uncertainty, data mining and statistical inference. Journal of the Royal Statistical Society A; 158(3), pag 419-466.

Geer, F.C. van (1988) Verlagingsberekening met transfermodellen rond de winplaats Spannenburg; H_2O; 21(16), pag 451-454.

Gehrels, J.C. (1995) Niet-stationaire grondwatermodellering van de Veluwe. Een studie naar de invloed van de grondwaterwinning, inpoldering en verloofing op de grondwaterstand sinds 1951; Rapport Vrije Universiteit, Amsterdam.

Hipel, K.W. en A.I. McLeod (1994) Time series modelling of water resources and environmental systems; Elsevier, Amsterdam.

Knotters, M. en M.F.P. Bierkens (1999) Tijdreeksmodellen voor de grondwaterstand: een kijkje in de black box; Stromingen; 5(3), pag 35-49.

Knotters, M. en J.G. de Gooijer (1999) TARSO modelling of water table depths; Water Resources Research; 35(3), pag 695-705.

Knotters, M., T. Hoogland en M. Pleijter (2011) Actualisatie van de grondwatertrappenkaart van holoceen Nederland. Opzet van het onderzoek; Wageningen, Alterra-rapport 2280.

McLeod, A.I. en K.W. Hipel (1978) Simulation procedures for Box-Jenkins models; Water Resources Research; 14(5), pag 969-975.

Oreskes, N., K. Shrader-Frechette en K. Belitz (1994) Verification, validation, and confirmation of numerical models in the earth sciences; Science; 263(5147), pag 641-646.

Rolf, H.L.M. (1989) Verlaging van de grondwaterstanden in Nederland. Analyse periode 1950-1986; Rapport DGV-TNO, Delft.