Hoezo plausibel? Plausibiliteit in de hydrologie

MARTIN KNOTTERS, PAUL BAGGELAAR, EIT VAN DER MEULEN EN DENNIS WALVOORT

'Plausibel' is een aanprijzing die met enige regelmaat wordt gegeven aan uitkomsten van hydrologische studies. Het schept vertrouwen als deskundigen onderzoeksresultaten 'plausibel' noemen. Maar wat bedoelen ze daar precies mee en bedoelen ze er allemaal hetzelfde mee? Is plausibiliteit objectief en reproduceerbaar vast te stellen? Wat is de rol van consensus bij de beoordeling van plausibiliteit? En als een deskundige een meetwaarde niet plausibel vindt, ligt dat dan aan de meting of aan de deskundigheid?

Essay

Inleiding

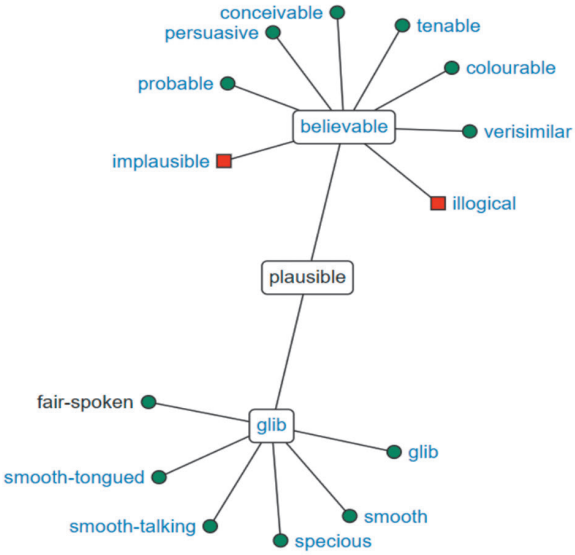
Dit essay is de derde publicatie in de serie "Hoezo...?", over veelgebruikte aanprijzingen van onderzoeksresultaten. Eerder verschenen de artikelen "Hoezo representatief? Over de betekenissen van 'representatief' in de KRW-literatuur" (Knotters e.a., 2009) en "Hoezo significant? Over het effect van een ingreep op de grondwaterstand" (Knotters e.a., 2020). Termen als 'representatief', 'significant' en 'plausibel' scheppen vertrouwen in onderzoeksresultaten. Als die onderzoeksresultaten dienen als basis voor beslissingen, dan is het goed om te weten wat nu precies bedoeld wordt met 'representatief', 'significant' en nu dus ook met 'plausibel'. In dit essay zetten we de verschillende betekenissen van 'plausibel' op een rijtje en brengen we het gebruik van 'plausibel' in de Nederlandse hydrologie en het waterbeheer in kaart. Ten slotte geven we onze gedachten over kwantitatieve, reproduceerbare alternatieven voor de beoordeling van plausibiliteit en beargumenteren we waarom de tijd voor deze alternatieven rijp is.

Wat betekent 'plausibel'?

Glette-Iversen e.a. (2022) zetten recent 17 definities en interpretaties van het woord 'plausibiliteit' uit de literatuur op een rijtje. Daarbij onderscheidden zij twee categorieën: plausibiliteit om de mate van waarschijnlijkheid uit te drukken en plausibiliteit om de sterkte van kennis aan te geven. De 17 definities en interpretaties van plausibiliteit vielen in één van deze categorieën of in beide. Voorbeelden van plausibiliteit als maat voor waarschijnlijkheid zijn de definitie van Morgan en Keith (2008): "*Plausibility is a synonym for relative subjective probability*", en van Nordmann (2013): "*Plausibility is equal to serious possibility*". Nussbaum (2011) definieert plausibiliteit als maat voor de sterkte van kennis: "*A proposition is plausible if it is reasonable to accept the proposition*". Een definitie van plausibiliteit die zowel op waarschijnlijkheid als sterkte van kennis betrekking heeft geeft de Oxford English Dictionary: "*The quality in an argument, statement, etc., of seeming reasonable or probable; appearance of reasonableness; believability, credibility*".

Van Dale's *online* woordenboek definieert 'plausibel' als 'aannemelijk, geloofwaardig' en sluit daarmee aan bij de definities die zowel aannemelijkheid als sterkte van kennis uitdrukken. Van Dale's etymologisch woordenboek (Van Veen en Van der Sijs, 1997) dateert de introductie van 'plausibel' in de Nederlandse taal op 1830 als leenwoord van het Franse *plausible*, beide met de betekenis 'aannemelijk'. Het Franse *plausible* komt van het Latijnse *plausibilis*, dat 'in de gunst staand' of eigenlijk 'beklapbaar' betekent, van *plaudere* (verleden deelwoord *plausum*): klappen (vergelijk 'applaudisseren'), met daaraan toegevoegd *-(b)ilis*, dat staat voor '-baar'.

De thesaurus van *The Free Dictionary* (*https://www.freethesaurus.com/plausible*, 3 maart 2022, 16.51 u) geeft een beeld van de synoniemen en antoniemen van *plausible* (zie afbeelding 1). Opmerkelijk is dat, althans in het Engelse taalgebied, 'plausibel' ook in verband kan worden gebracht met *glib*. Dit kan zowel de positieve connotaties van vlot en welbespraakt hebben, als de negatieve van rad van tong, mooipraterij en 'schoonschijnend' (*specious*).

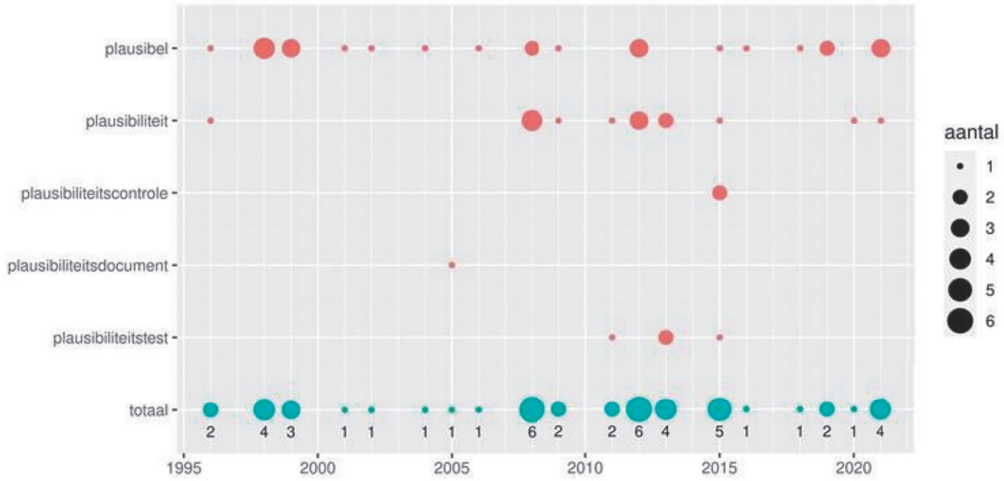


Afbeelding 1: *Synoniemen (groene bolletjes) en antoniemen (rode vierkantjes) voor 'plausible'. Bron: https://www.freethesaurus.com/plausible, 3 maart 2022, 16.51 u*

Samenvattend is 'plausibel' een aanprijzing met de betekenissen 'aannemelijk' en 'geloofwaardig' en ontbreekt een kwantitatieve, reproduceerbare onderbouwing.

'Plausibel' in de hydrologie

Afbeelding 2 laat de frequentie zien waarmee 'plausibel' voorkomt in Stromingen: we zien magere jaren en vette, zoals momenteel. We onderzochten ook de frequentie waarmee 'plausible' en 'plausibility' voorkomen in de titels, trefwoorden en samenvattingen van de Journal of Hydrology, Water Resources Research en Hydrology and Earth System Science (afbeelding 3). Daaruit blijkt een toenemend gebruik van deze termen.



Afbeelding 2: Frequentie van 'plausibel' in Stromingen

Afbeelding 3: Frequentie van 'plausible' en 'plausibility' in titels, trefwoorden en samenvattingen van de tijdschriften Journal of Hydrology, Water Resources Research en Hydrology and Earth System Science

'Plausibel' wordt gebruikt met betrekking tot zowel modellen en modeluitkomsten (schattingen, voorspellingen), als waarnemingen. Een voorbeeld van het gebruik van 'plausibel' in relatie tot modellen en modeluitkomsten is het volgende citaat uit een persbericht dat LTO op 19 oktober 2020 deed uitgaan naar aanleiding van het vrijgeven van Waterwijzer Landbouw (WWL) voor toepassing in de praktijk: "*De onderzoekers concluderen dat de uitkomsten van WWL op verschillende fronten niet plausibel en niet consistent zijn met algemeen aanvaarde hydrologische kennis*". Het persbericht verwijst naar een studie van Van Bakel en Hoogewoud (2020), waarin de volgende betekenis aan 'plausibel' wordt gegeven: in overeenstemming met een gangbare opvatting (blz. 33), met praktijkkennis (blz. 34), met de HELP-tabel (blz. 35) of met algemeen aanvaarde uitgangspunten (blz. 37). Het is toe te juichen dat vermeld wordt wat met 'plausibel' wordt bedoeld. Toch rijzen er enkele vragen bij ons: 1) Ligt het oordeel 'niet plausibel' aan WWL, aan de 'algemeen aanvaarde hydrologische kennis' of aan beide?; 2) Staat 'algemeen aanvaarde kennis' vast als een rots, of is voortschrijdend inzicht mogelijk en kunnen de uitkomsten van WWL hier misschien aanleiding toe geven?; en 3) Zeggen plausibiliteit en consistentie met algemeen aanvaarde hydrologische kennis genoeg over de validiteit van WWL of is daar meer voor nodig? De wetenschappelijke adviescommissie NHI/NWM concludeert dat de WWL op dit moment '*fit-for-purpose*' is mits aan een aantal randvoorwaarden voldaan wordt, waarvan de belangrijkste een op plausibiliteit en consistentie gebaseerd draagvlak is (Peerboom e.a., 2022). In tegenstelling tot Van Bakel en Hoogewoud (2020) geven zij echter niet aan wat met plausibiliteit wordt bedoeld. Wij bepleiten verderop in dit essay toepassing van kwantitatieve en reproduceerbare alternatieven voor plausibiliteitsbeoordeling, of in elk geval het hanteren van een duidelijke definitie van plausibiliteit.

Het IPCC onderbouwt de verwoording van onzekerheid in termen van *confidence* en *likelihood* met *calibrated language* (IPCC, 2010). Bij deze onderbouwing wordt de niet nader onderbouwde term '*plausible*' gebruikt: "*Consider all plausible sources of uncertainty*". Niet-plausibele bronnen van onzekerheid worden dus buiten beschouwing gelaten, maar wat plausibel is en waar de grens ligt met niet-plausibel is niet aangegeven. Dit is een gemis, want het ontbreken van een definitie van 'plausibel' in IPCC (2010) maakt de onderbouwing van de verschillende gradaties van *confidence* en *likelihood* minder sterk.

Alleen plausibel is niet genoeg

De handleiding Tijdreeksanalyse (STOWA, 2021) vermeldt beoordeling van tijdreeksmodellen op basis van wat hydrologisch plausibel is (blz. 171-172), maar bevat ook de volgende kanttekening (blz. 58): "Door sommige analisten wordt een model ook wel gekwalificeerd als plausibel of niet-plausibel, afhankelijk van het wel of niet aansluiten op fysisch-hydrologische logica. De term plausibel kan echter een té positief oordeel over het model uitstralen, aangezien deze een combinatie van de volgende eigenschappen suggereert: aannemelijk, aanvaardbaar, redelijk, waarschijnlijk, begrijpelijk, logisch, overtuigend, geloofwaardig of steekhoudend. Dergelijke positieve kwalificaties gaan veel verder dan te rechtvaardigen is uit een vaststelling dat het model niet in tegenspraak is met fysisch-hydrologische logica." Een gesloten waterbalans is bijvoorbeeld

een fysische vereiste en is plausibel, maar dat wil nog niet zeggen dat de balanstermen correct zijn. Die kunnen onnauwkeurig zijn, zoals bijvoorbeeld uit een validatie kan blijken. En hoewel het plausibel is dat oppervlaktewater van hoog naar laag stroomt, hoeven de debieten nog niet juist te zijn en afwijken van de werkelijke debieten.

Vanaf de tweede helft van de jaren tachtig zijn Nederlandse hydrologen tijdreeksanalyse gaan toepassen, in eerste instantie vooral als hulpmiddel bij onderzoek naar de invloed van een grondwaterwinning op de grondwaterstand. Tot ongeveer 2005 betrof dit de Box-Jenkins-methode (Box and Jenkins, 1970) en dan in het bijzonder het transferruismodel (hierna aangeduid als BJTR-model). Dat model stelt namelijk in staat eventuele statistische relaties in beeld te brengen tussen een hydrologische variabele, zoals de grondwaterstand, en mogelijke invloedsfactoren, zoals neerslag, verdamping en grondwaterwinning. Daarbij levert het BJTR-model per invloedsfactor niet alleen een raming van zijn evenwichtsrelatie met de grondwaterstand, maar ook zijn onzekerheid, zoals uitgedrukt in het 95%-betrouwbaarheidsinterval. En indien gewenst kunnen die relaties dan ook worden benut, bijvoorbeeld om de grondwaterstand te beheren, of te voorspellen, waarbij behalve de voorspellingen ook hun onzekerheden worden meegegeven, zoals met het 95%-voorspelinterval.

Om een BJTR-model te kunnen ontwikkelen dienen de tijdreeksen van de te modelleren reeks en die van de invoerreeksen equidistant te zijn en ook alle parallel in de tijd te lopen. Het ontwikkelen geschiedt in drie fasen: 1) identificeren van het model (langs empirische weg), 2) schatten van de modelparameters en 3) verifiëren van het model. Als in de verificatiefase blijkt dat het BJTR-model voldoet aan een aantal randvoorwaarden, waaronder dat het modelresidu een aselecte trekking is uit een normale kansverdeling, kan er van worden uitgegaan dat: 1) de modelparameters zuiver zijn geschat, dat wil zeggen zonder systematische fout en 2) de dekkingsgraad van een meegeleverd betrouwbaarheidsinterval en/of voorspelinterval gelijk is aan de door de analist ingestelde dekkingsgraad (zoals 95%). De meerwaarde van deze eigenschappen van de resultaten is enorm. Ze stellen namelijk in staat objectief vast te stellen of een geschatte statistische relatie wel of niet statistisch significant afwijkt van nul en daarmee of er überhaupt wel sprake is van een relatie, gegeven de modelonzekerheden. Dit maakt de resultaten van een BJTR-model bijvoorbeeld geschikt als bewijsmateriaal bij een juridisch dispuut.

Hydrologische toepassingen van het BJTR-model bleven echter bijna twee decennia beperkt tot een selecte groep specialisten. Wellicht speelde mee dat geschikte software kostbaar was, of zelf ontwikkeld moest worden, maar aannemelijker is dat velen afzagen van toepassing omdat het kennis van statistiek vereist en die ontbreekt bij veel Nederlandse hydrologen. Ter vergelijking: vanaf begin jaren tachtig is de BJTR-methode wel breed toegepast in de econometrie, een discipline waar in de opleiding veel aandacht wordt besteed aan statistiek.

Pas vanaf 2005 begon het toepassen van tijdreeksanalyse door Nederlandse hydrologen toe te nemen, doordat een vorm van tijdreeksanalyse beschikbaar

kwam die geen statistische kennis vereiste van de analist, waar betaalbare software voor beschikbaar was en die geen equidistante en parallelle tijdreeksen vergde. Dat laatste was juist in die periode een voordeel, omdat geleidelijk het omschakelen begon van halfmaandelijks handmeten van de grondwaterstand, naar hoogfrequent meten met een automatische logger. Het betrof uitvoeringen van de zogenaamde PIRFICT-methode[1] (zie bijvoorbeeld Von Asmuth e.a., 2002), in eerste instantie met het gebruiksvriendelijke programma Menyanthes en recenter ook met PASTAS, een open-source-verzameling Pythonroutines (Collenteur e.a., 2019). Dit alles leidde er toe dat inmiddels een groot aantal Nederlandse en buitenlandse hydrologen tijdreeksanalyse toepast met de PIRFICT-methode.

Het is uiteraard een goede zaak dat tijdreeksanalyse een populair instrument is geworden voor veel hydrologen en niet meer louter voorbehouden is aan een select groepje deskundigen. Maar toch moet daar ook de volgende kanttekening bij worden geplaatst. Het blijkt dat de meesten van de nieuwe lichting tijdreeksanalisten een PIRFICT-model vooral beoordelen op de pasvorm (*goodness-of-fit*), uitgedrukt in de maat EVP (Explained Variance Percentage). Die vertegenwoordigt het percentage van de variantie van de tijdreeks dat wordt verklaard door het deterministische deel van het model, zoals het deel dat wordt verklaard door neerslag en verdamping, als dat de enige invoerreeksen zijn. Een model wordt door PIRFICT-analisten doorgaans 'plausibeler' geacht naarmate de EVP hoger is en velen hanteren daarbij als ondergrens 70% (ook al is die grens nooit onderbouwd). Maar nog afgezien van het feit dat de EVP vaak verkeerd wordt berekend[2], is deze ongeschikt als kwaliteitsmaat. Een hoge EVP geeft immers nog geen garantie dat de modelparameters zuiver zijn geschat (dat wil zeggen zonder systematische fout) en dat de dekkingsgraden van betrouwbaarheids- en voorspelintervallen overeenstemmen met de ingestelde dekkingsgraden. Een dergelijke garantie vereist onder meer dat het modelresidu (aangeduid als 'ruis' in de PIRFICT-terminologie) normaal verdeelde witte ruis vormt en dat zal dan eerst moeten worden geverifieerd aan de hand van de combinatie van het autocorrelogram en het partieel autocorrelogram van dat residu. Maar in de praktijk blijkt die verificatie door veel PIRFICT-analisten niet te worden uitgevoerd. Blijkbaar geeft een voldoende hoge EVP ze al voldoende vertrouwen dat de modelresultaten bruikbaar zijn, maar dat vertrouwen is in dergelijke gevallen onterecht. Het wijst op onbekendheid met de randvoorwaarden die gelden voor gebruik van een tijdreeksmodel. Baggelaar en Van der Meulen (2020) toonden met een eenvoudig simulatie-experiment aan dat bij niet voldoen aan die randvoorwaarden de dekkingsgraad van een betrouwbaarheidsinterval tot twee of driemaal lager kan zijn dan ingesteld. De kans op onterecht detecteren van een statistisch significante relatie tussen een mogelijke invloedsfactor en een grondwaterstand kan dan oplopen tot meer dan 50%.

1 PIRFICT is het acronym van Predefined Impuls-Response Functions In Continuous Time.

2 Deze wordt meestal berekend als 100% min 100% maal de ratio van de variantie van de ruis (aangeduid als 'residu' in de PIRFICT-terminologie) en de variantie van de grondwaterstandreeks, maar daarbij wordt vergeten de covariantie van het deterministische model en de ruis te verdisconteren.

Met andere woorden: *goodness-of-fit* zegt nog niets over *fitness-for-use.* De pasvorm dient slechts een rol te gaan spelen als er meerdere modellen zijn die blijken te voldoen aan de randvoorwaarden en er één model moet worden geselecteerd. Het model met bijvoorbeeld de hoogste verklaarde variantie komt dan het meest in aanmerking, eventueel in combinatie met een *penalty* voor modelcomplexiteit, zoals bij Akaike's en Bayes' Information Criterion (AIC en BIC). Dit is in de BJTR-methode gebruikelijk. Ook zou kruisvalidatie kunnen worden toegepast bij modelselectie, wat de mogelijkheid biedt om het validatiecriterium af te stemmen op de toepassing van het model, zodat het model wordt geselecteerd dat het meest *fit-for-use* is (Knotters, 2012).

Plausibiliteitstoetsen bij kwaliteitscontroles

Hier en daar bestaan gevorderde plannen om met loggers gemeten grondwaterstanden te valideren met een tijdreeksmodel voor die data. Metingen die buiten het voorspelinterval liggen worden daarbij minder plausibel geacht. Het gebruik van modellen bij de controle van metingen vinden we echter een riskante ontwikkeling, omdat het indruist tegen de wetenschappelijke methode: je leert niet van de waarnemingen. Met data kan een model worden gevalideerd, maar als met een model data worden gevalideerd, ontstaan er allerlei risico's, zoals op stilstand van onze kennis. Als tijdreeksmodellen worden gebruikt bij kwaliteitscontrole van meetreeksen, dan moeten die niet al te rigoureus de voorkeur van de twijfel krijgen als weergave van de werkelijkheid, want het gevolg is dat data worden afgekeurd die ons juist over de werkelijkheid informeren. De hydrologische processen die tijdreeksen genereren zijn niet te vergelijken met industriële processen, die volledig door de mens zijn ontworpen en gereguleerd en waarvan alle ins en outs volledig bekend zijn (gesloten systemen). Afwijkingen van het model duiden dan op productiefouten (of storingen/slijtage in de processtroom). Deze expertise kun je niet zomaar toepassen op natuurlijke systemen.

Een voorbeeld van 'plausibel' in relatie tot meetgegevens geeft het Protocol voor datakwaliteitscontrole (QC), versie 2.0 (Platform meetnetbeheerders grondwaterkwantiteit, 2018), waarin het onderdeel 'Plausibiliteit' zich richt op de vraag hoe waarschijnlijk of plausibel (meet)gegevens zijn gezien onder meer vakinhoudelijke kennis. Daarbij vragen wij ons af of niet-plausibele metingen een lage kwaliteitsbeoordeling moeten krijgen of dat de vakinhoudelijke kennis wellicht verrijkt kan worden met een nieuw inzicht naar aanleiding van niet-plausibele metingen. Wij juichen toe dat het tijdschrift Stromingen een rubriek 'Rare reeksen' kent, die ruimte biedt om hydrologische inzichten te verdiepen op basis waarnemingen. Als je de rubriek doorneemt, dan blijken die rare reeksen ons goed aan het denken te zetten.

Plausibiliteit, groepsdenken en vertekening

In de hydrologie en het waterbeheer in Nederland wordt veel waarde gehecht aan consensus en draagvlak. Beide zijn het resultaat van groepsprocessen, zoals plaatsvinden in begeleidingscommissies, klankbord- en stuurgroepen en wat dies meer zij. Groepsprocessen kunnen echter leiden tot vertekening, zoals te veel vertrouwen en een onderschatting van onzekerheid. Dit wordt ook wel

groupthink genoemd. Baron (2000, blz. 216) schrijft hierover: "*A group consensus sometimes seems much more obviously "right" than the same conclusion reached by an individual, even though the members of the group are all alike in sharing the characteristics that lead to the consensus*". Ook het IPCC (2010) waarschuwt voor *groupthink*: "*Be aware of a tendency for a group to converge on an expressed view and become overconfident in it*", waarbij gewezen wordt op het gevaar van *anchoring*: nieuwe schattingen worden meer bepaald door eerdere schattingen dan gerechtvaardigd is. Conformisme kan een rol spelen doordat groepsleden met afwijkende inzichten zich aanpassen aan de meerderheid, de 'algemeen aanvaarde kennis'. Groepsdenken over plausibiliteit kan ook vertekend raken door het collectief optreden van *confirmation bias*. Daarbij wordt meer gewicht gegeven aan informatie die je hypothese bevestigt dan aan informatie die je hypothese ontkracht: zie je wel, dat zei ik ook al. Ook kan bij plausibiliteitsbeoordeling, zowel op individueel als groepsniveau, sprake zijn van *narrative bias*: de tendens die mensen van nature hebben om patronen of structuren te zien en gebeurtenissen te kunnen verklaren. Gebeurtenissen worden in een verhaal geplaatst, ongeacht of deze gebeurtenissen het verhaal ondersteunen of meer toevallig zijn (Betsch e.a., 2015). Voor een overzicht van de vertekeningen die bij beoordelingen onder onzekerheid kunnen optreden verwijzen wij naar het Science-artikel van Tversky en Kahneman (1974) en naar het boek 'Ons feilbare denken/Thinking, fast and slow' (Kahneman, 2011).

Plausibiliteit in deze tijd

We zien de frequentie waarmee in de hydrologie over plausibiliteit wordt gesproken toenemen. We zien echter ook drie ontwikkelingen in de samenleving die vragen om een kwantitatieve onderbouwing van beslissingen, die niet geboden kan worden met alleen een plausibiliteitsoordeel.

De eerste ontwikkeling komt overwaaien uit de Angelsaksische wereld: de toenemende claimcultuur. Het ontstaan van droogteschade.nl is hier een voorbeeld van. In rechtszaken zal aan het kwalitatieve plausibiliteitsoordeel van een getuige-deskundige minder gewicht worden gegeven dan aan kwantitatieve informatie over bijvoorbeeld de omvang van een effect van een ingreep in de waterhuishouding.

De tweede ontwikkeling is dat burgers en boeren steeds meer zelf op onderzoek uitgaan en data verzamelen. Internet en basisregistraties maken dit onder meer mogelijk. De overheid moet hierdoor haar beslissingen kwantitatief beter onderbouwen dan met alleen een plausibiliteitsoordeel mogelijk is. Een voorbeeld van deze ontwikkeling is de oprichting van Stichting Agrifacts. In de discussie rond de stikstofcrisis vragen boeren bijvoorbeeld om onderbouwing van het beleid met metingen.

De derde ontwikkeling is de toenemende aandacht voor doelmatigheid en doeltreffendheid van beleid, wat is af te lezen uit de komst van rekenkamers en rekenkamercommissies bij overheden. Het bepalen van doelmatigheid en doeltreffendheid vraagt om kwantitatieve informatie, bijvoorbeeld over risico's (kans maal effect).

Het kan ook anders

Voor plausibiliteitstoetsen van modellen en data bestaan kwantitatieve en reproduceerbare alternatieven. In het STOWA-handboek Good Modelling Practice (STOWA, 1999) komt het woord 'plausibel' niet voor en het bevat ook niet de aanbeveling om een plausibiliteitstoets uit te voeren. Het vermeldt daarentegen wel validatie en verificatie. Validatie is het "vergelijken van modeluitvoer met een onafhankelijke (dat wil zeggen niet in de kalibratie gebruikte) set meetgegevens, teneinde te kunnen vaststellen of het model 'goed' is (of het concept goed is, het model het verleden kan reproduceren met de vereiste nauwkeurigheid en of het model geschikt is om alle vragen te beantwoorden". Verificatie is volgens het STOWA-handboek de "controle of het wiskundige model correct geïmplementeerd is in het computerprogramma".

Zoals eerder opgemerkt, wordt bij tijdreeksmodellering de term 'verificatie' gebruikt als vertaling van *diagnostic checking*: controleren of aan de modelveronderstellingen wordt voldaan. De uitkomst hiervan zegt meer over de bruikbaarheid van een model dan een plausibiliteitstoets waarbij de *goodness-of-fit* leidend is.

Bij kwaliteitscontrole van data zijn reproduceerbare alternatieven voor plausibiliteitstoetsen beschikbaar, zoals hypothesetoetsen (Neyman-Pearson), waarbij de risico's van ten onrechte goed- en afkeuren kunnen worden beheerst (zie Knotters e.a. (2020) voor het verschil tussen hypothese- en significantietoetsen). Ook kan voor een niet-statistische controle worden gekozen, die echter wel reproduceerbaar is, namelijk toepassing van *fuzzy logic* (Walvoort e.a., 2013). Hierbij kunnen kwaliteitsgradaties worden gedefinieerd op basis van expertbevraging (*expert elicitation*), waarbij specialistische statistische expertise niet nodig is. Voorwaarde is wel dat *expert elicitation* deskundig wordt uitgevoerd, om vertekeningen door groepsdenken te voorkomen. Hier is methodiek voor, zie bijvoorbeeld Hemming e.a. (2018).

Beslissingen in het waterbeheer zouden meer kunnen worden gebaseerd op de uitkomsten van (statistische) beslismodellen, in plaats van op deskundigenoordeel over plausibiliteit. Dit vraagt wel dat statistici in een vroeger stadium bij projecten worden betrokken dan vaak gebruikelijk is: hun kennis kan goed van pas komen bij het definiëren van het beslisprobleem.

We kunnen natuurlijk ook doorgaan met plausibiliteitsbeoordeling, maar dan met een heldere definitie van wat we onder 'plausibel' verstaan, *calibrated language* dus. Daarbij zouden we een voorbeeld kunnen nemen aan het populair wetenschappelijke televisieprogramma Myth Busters (Discovery Channel), *https://en.wikipedia.org/wiki/MythBusters#Plausible*, waarin mythes worden ontkracht, bevestigd of plausibel geacht op basis van nauwkeurig omschreven criteria.

Literatuur

Asmuth, J.R. von, M.F.P. Bierkens en K. Maas (2002) Transfer function noise modelling in continuous time using predefined impulse response functions; in: *Water Resources Research*, Volume 38, Issue 12, pag 23_1 - 23_12.

Baggelaar, P.K. en E.C.J. van der Meulen (2020) Naar betere tijdreeksmodellering met Pastas; AMO en PB Icastat, 28 april 2020.

Bakel, P.J.T. van en J.C. Hoogewoud (2020) Landsdekkende vergelijking HELP-methode met WaterWijzer Landbouw (WWL) voor de berekende droogte- en natschade van grasland. Een verkennende studie naar de toepasbaarheid van de WaterWijzer Landbouw, uitgevoerd in opdracht van Droogteschade B.V. De Bakelse Stroom/Advies in Water.

Baron, J. (2000) Thinking and deciding; Third Edition; Cambridge University Press.

Betsch, C., N. Haase, F. Renkewitz en Ph. Schmid (2015) The narrative bias revisited: what drives the biasing influence of narrative information on risk perceptions?; in: *Judgment and Decision Making,* vol 10(3), pag 241-264.

Box, G.E.P. en J.M. Jenkins (1970) Time series analysis, forecasting and control. Holden-Day, San Francisco.

Collenteur, R.A., M. Bakker, R. Caljé, S.A. Klop en F. Schaars (2019) Pastas: open source software for the analysis of groundwater time series; in: *Groundwater* doi: 10.1111/gwat.12925.

Glette-Iversen, I., T. Aven en R. Flage (2022) The concept of plausibility in a risk analysis context: Review and clarifications of defining ideas and interpretations; in: *Safety Science*, vol 147, 106535.

Hemming, V., M.A. Burgman, A.M. Hanea, M.F. McBride en B.C. Wintle (2018) A practical guide to structured expert elicitation using the IDEA protocol; in: *Methods in Ecology and Evolution*, vol 9, pag 169-180.

IPCC (2010) IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. IPCC Cross-Working Group Meeting on Consistent Treatment of Uncertainties, Jasper Ridge, CA, USA, 6-7 July 2010.

Kahneman, D. (2011) Ons feilbare denken/Thinking, fast and slow. Business Contact, Amsterdam/Antwerpen.

Knotters, M. (2012) Validatie van tijdreeksmodellen voor de grondwaterstand; in: *Stromingen*, vol 18(2), pag 31-42.

Knotters, M., D.J. Brus en J.J. de Gruijter (2009) Hoezo representatief? Over de betekenissen van 'representatief' in de KRW-literatuur; in: *Stromingen*, vol 15(1), pag 3-14.

Knotters, M., P.K. Baggelaar en E. van der Meulen (2020) Hoezo significant? Over het effect van een ingreep op de grondwaterstand; in: *Stromingen*, vol 26(3), pag 37-49.

Nordmann, A. (2013) (Im)Plausibility; in: *International Journal of Foresight and Innovation Policy*, vol 9(2/3/4), pag 125.

Nussbaum, E.M. (2011) Argumentation, dialogue theory, and probability modeling: Alternative frameworks for argumentation research in education; in: *Educational Psychologist*, vol 46(2), pag 84-106.

Peerboom, J., B. van den Hurk en H. van Waveren (2022) WWL: 'fit-for-purpose' of niet? Wetenschappelijke Adviescommissie NHI/NWM beoordeelt Waterwijzer Landbouw; in: Stromingen, vol 28(1), pag 19-23.

Platform meetnetbeheerders grondwaterkwantiteit (2018) Protocol voor datakwaliteitscontrole (QC) versie 2.0. Kwaliteitsborging grondwaterstands- en stijghoogtegegevens; Platform meetnetbeheerders grondwaterkwantiteit van de gezamenlijke provincies, KWR Watercycle Research Institute.
STOWA (1999) Vloeiend modelleren in het waterbeheer. Handboek Good Modelling Practice; STOWA-rapport 99-05, Rijkswaterstaat-RIZA-rapport 99.036.
STOWA (2021) Handleiding tijdreeksanalyse; STOWA, Rapport 32.
Tversky, A. en D. Kahneman (1974) Judgment under uncertainty: heuristics and biases; in: Science, vol 185, pag 1124-1131.
Veen, P.A.F. van en N. van der Sijs (1997) Etymologisch woordenboek. De herkomst van onze woorden; Van Dale Lexicografie, Utrecht/Antwerpen.
Walvoort, D., M. Knotters en T. Hoogland (2013) Een tool voor controle van hoogfrequente grondwaterstandsreeksen; in: *Stromingen*, vol 19(3,4), pag 49-61.

Summary What do you mean 'plausible'? Plausibility in hydrology

Hydrological research outcomes are increasingly applauded by calling them 'plausible', and model predictions as well as observed values of hydrological variables are judged on 'plausibility'. It is not always entirely clear, however, what exactly is meant with 'plausible'. Origin from the Latin 'plausibilis' (able to be applauded) the intents of 'plausible' deviate from an indication of subjective likelihood to a measure of knowledge strength. However, plausibility judgments are not quantitatively and consistently reproducible. In hydrological time series modelling a tendency to select models on the basis of plausibility with a focus on goodness-of-fit is observed, rather than on diagnostic checks to verify whether the model assumptions are answered, which is a prerequisite for use in risk assessment, i.e. fitness-for-use. In quality control of observed time series of groundwater levels a tendency to test plausibility of observed values by comparing with time series model predictions is noticed. This practice might hinder the contribution of observations to the development of hydrological knowledge. Judgments on plausibility can be biased as a result of groupthink, since consensus among experts is often strived for in Dutch hydrological and water management practices. Besides this, heuristics to assess probabilities might lead to bias in individual plausibility judgments. Three actual trends urge for more quantitatively substantiated assessment of hydrological research outcomes than can be provided with plausibility judgments: i) an increasingly juridification in society, ii) citizens and farmers collecting data themselves and asking governments to support decision making by data, and iii) an increased emphasis to efficiency and efficacy of policy. Alternative to plausibility judgments we suggest: validation of model outcomes using independent reference data, verification of model assumptions in time series modelling by diagnostic checking of the residuals, quality control by hypothesis testing (Neyman-Pearson) or by fuzzy logic supported with prudent expert elicitation, and application of statistical models for decision making in water management. However, if the use of plausibility judgments in hydrology is continued, then a clear definition and quantitative basis to 'plausible' should be created, i.e. 'calibrated language'.

Auteurs

MARTIN KNOTTERS
Wageningen University & Research
martin.knotters@wur.nl

PAUL K. BAGGELAAR
PB Icastat
paul.baggelaar@planet.nl

EIT C.J. VAN DER MEULEN
Amo
amo@amo-nl.com

DENNIS J.J. WALVOORT
Wageningen University & Research
dennis.walvoort@wur.nl