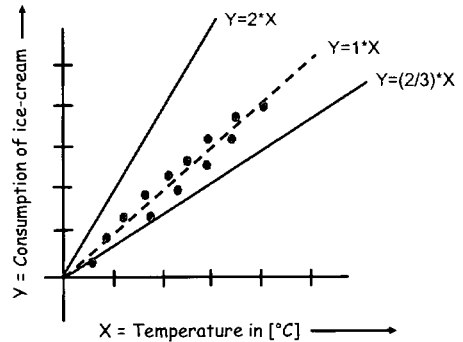


Aquapodium promovendis

4 Februari 2005, ik ontvang een e-mail van Joost Heijkers met een uitnodiging om een stukje te schrijven over mijn promotieonderzoek voor het blad *Stromingen* van de Nederlandse Hydrologische Vereniging. Ik ben op dat moment in Tucson, Arizona en bezig met de laatste onderhandelingen voor mijn Director's gesubsidieerde postdoc positie aan het Los Alamos National Laboratory (LANL) in New Mexico, USA. Deze baan geeft me volledige vrijheid in mijn onderzoek binnen de discipline van aardwetenschappen. Ik heb net de verplichte veiligheidstraining aan LANL afgerond en zit nu, inmiddels al weer acht weken later, en ongeveer een week voor de deadline, achter mijn computer in Los Alamos terwijl buiten de sneeuw gestaag valt.

Het promotieonderzoek dat ik eind Juni 2004 heb afgerond, heeft zich gericht op het ontwikkelen van parameteroptimalisatie-technieken voor het kalibreren van hydrologische modellen. Dit onderzoek heb ik uitgevoerd aan de Universiteit van Amsterdam binnen de groep Computationale Bio- en Fysische Geografie (CBPG) van Willem Bouten. Veel van de parameters in hydrologische modellen beschrijven geaggregeerde ruimtelijke en temporele processen, en zijn daardoor veelal niet op een onafhankelijke manier vast te stellen. Neem als voorbeeld de doorlatendheid van een polder. Hoe stel je die vast? Je kunt een aantal ringmonsters nemen en dan in een bodemfysisch laboratorium van elk van deze monsters de doorlatendheid bepalen. Maar hoe aggregere je de verschillende uitkomsten van deze experimenten tot een gemiddelde doorlatendheid, als de verschillende processen die optreden hoog niet-lineair zijn? Invers modelleren of parameteroptimalisatie is hierbij een uitkomst.



Figuur 1: Simpel voorbeeld om inverse modellering te illustreren. De cirkels geven de metingen aan, terwijl de lijnen verschillende modelfits beschrijven.

Laten we beginnen met een simpel voorbeeld. In figuur 1 staat de ijsconsumptie (y -as) afgezet tegen de temperatuur (x -as). Dit betreft een set van fictieve metingen die op verschillende dagen is vastgesteld. Stel nu dat we een model door deze metingen zouden willen fitten, en dit model is een rechte lijn ($y = ax$), met één parameter, a . De optimale waarde van de parameter a kunnen we vrij aardig vast stellen door verschillende a -waarden te proberen, en de visuele uitkomst tussen modelfit en metingen te vergelijken. Met een semi-intuïtieve iteratieve methode kunnen we dan de optimale waarde van a vaststellen. In dit specifieke triviale voorbeeld is dat natuurlijk vrij eenvoudig. Maar hoe zit dat nu, als het onderliggende model (de lijn) niet één parameter bevat, maar bijvoorbeeld 10 of 20? Met een toenemend aantal parameters wil je natuurlijk niet handmatig aan de 'knoppen' blijven draaien. Een uitkomst zou een computergestuurd algoritme zijn, dat de parameteroptimalisatie geheel automatisch uitvoert. Een minstens net zo belangrijke vraag is hoe het zit met de bijbehorende onzekerheid van onze uiteindelijke parameterschatting? Gegeven fouten in de meting, en het onderliggende model, kunnen we natuurlijk niet zonder mee aan nemen dat onze optimale schatting niet onderhevig is aan

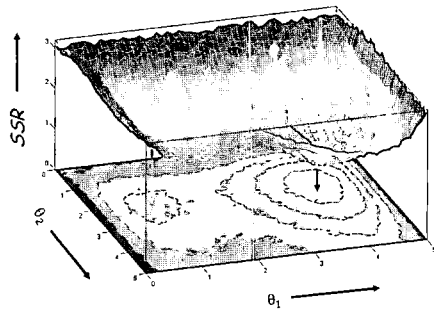
fouten. Welke meting van (x,y) bevat überhaupt de meeste informatie om onze modelparameter vast te stellen? Deze informatie zou erg helpen bij het ontwikkelen van optimale experimentele meetstrategieën. Met dit soort vragen heb ik me intensief bezig gehouden in mijn promotie-onderzoek.

Vanaf de jaren '50 zijn er verschillende optimalisatiealgoritmen gepubliceerd voor automatische kalibratie van modelparameters. Voorbeelden van dergelijke algoritmen die wijdverspreide implementatie hebben gevonden in verschillende vakgebieden zijn Gauss-Newton, Levenberg-Marquardt (LM) en de Simplex-methode. Deze technieken staan bekend onder lokale zoekmethoden omdat ze beginnen met een (veelal aselekt gekozen) initiële parametercombinatie, en zoeken voor iteratieve verbetering vanuit dit startpunt door gebruik te maken van een aantal vaste algoritmische stappen. In elke stap wordt een geaggregeerde lengtemaat van het verschil tussen model en meting berekend, en de nieuwe parametercombinatie als nieuw startpunt geaccepteerd als dit verschil afneemt. Een lengtemaat die veel gebruikt wordt in hydrologisch modelleren is de Sum of Squared Errors (SSR):

$$SSR = \sum_{t=1}^n (y_t - \tilde{y}_t)^2$$

waarin t staat voor tijd, n het aantal waarnemingen, en $(y_t - \tilde{y}_t)$ het verschil tussen modelvoorspelling (\tilde{y}_t) en meting (y_t) is. Ter illustratie bevat figuur 2 een voorbeeld hoe het LM-algoritme convergeert naar een optimale parameterschatting voor een fictief tweedimensionaal optimalisatieprobleem, gegeven twee verschillende startpunten (cirkels) in de parameter ruimte. De sequentiële evolutie van de met LM geschatte 'optimale' parameterwaarde is

geïndiceerd met een donkere lijn. Zoals uit figuur 2 blijkt is de LM-methode computationeel erg efficiënt als het de kortste weg vindt in de parameter ruimte tot de best mogelijke iteratieve verbetering. Dit is niet verwonderlijk aangezien het algoritme direct gebruik maakt van de gradiënt van het response-oppervlak om de richting van grootste verbetering vast te stellen.



Figuur 2: Responseoppervlak voor een fictief twee-parameter-optimalisatieprobleem. De donkere lijnen stellen de evolutie voor van een Levenberg-Marquardt-optimalisatie.

Ondanks het feit dat lokale zoekmethoden snel convergeren naar een 'optimale' oplossing, hebben deze methoden ook een ogenschijnlijk belangrijke tekortkoming. Dit wordt duidelijk geïllustreerd in figuur 2. Indien het response-oppervlak meer dan één goed gedefinieerd minimum vertoont, iets wat typisch is voor hydrologische modellen, is de uiteindelijke oplossing sterk afhankelijk van de startlocatie. Hoewel dit probleem natuurlijk op te lossen is door een groot aantal optimalisaties uit te voeren met verschillende begincondities, zal dit in de praktijk computationeel erg inefficiënt zijn, omdat elk van deze optimalisaties volledig onafhankelijk opereert. Ook wordt het probleem van lokale minima met toenemende dimensionaliteit steeds ingrijpender, wat vraagt om een computationeel meer effectieve en robuuste methode.

Een betere optimalisatiestrategie zou zijn om simultaan verschillende lokale

zoekmethoden te draaien om deze onafhankelijk van elkaar, en in een verschillend deel van de parameterruimte, informatie over het response-oppervlak te laten inwinnen. Na enige tijd zouden deze verschillende initialisaties dan hun informatie moeten uitwisselen om vervolgens weer door te gaan met hun iteratieve verbetering. Periodieke uitwisseling van informatie gedurende korte bijeenkomsten zal convergentie naar het absolute optimum faciliteren. Dit voortschrijdende inzicht heeft geresulteerd in de ontwikkeling van globale optimalisatietechnieken. In tegenstelling tot lokale zoekmethoden, maken deze algoritmen gebruik van een populatie van parametersets, met verschillende startlocaties, om effectief en efficiënt de parameterruimte door te zoeken. Deze zoekmethoden zijn daardoor minder gevoelig voor lokale minima wat convergentie naar het globale optimum garandeert, maar wel ten koste van een toegenomen computationele tijd.

Een voorbeeld van een zoekmethode die voldoet aan de beschrijving van specialisatie gevolgd door informatieuitwisseling is het Shuffled Complex Evolution (SCE-UA) globale optimalisatiealgoritme, ontwikkeld aan de Universiteit van Arizona. De details van dit algoritme kunnen worden gevonden in Duan e.a. (1992). Dit algoritme kan stapsgewijs beschreven worden als:

- 1 Genereer een initiële parameterpopulatie en bereken de bijbehorende *SSR* van elk punt;
- 2 Sorteer de punten in toenemende *SSR* waarde;
- 3 Verdeel de punten onder in verschillende complexen;
- 4 Evolueer elk van deze complexen met behulp van de Simplex-methode (vervang slechtste punt in complex als Simplex-gegenereerde parametercombinatie beter is dan dit punt);
- 5 Na een aantal stappen in elk complex: gooi alle punten in de verschillende complexen weer op één hoop, en ga terug

naar stap (2).

Hoewel verschillende case studies hebben gerapporteerd dat dit algoritme uitstekende convergentie-eigenschappen heeft voor hoogdimensionale problemen, houdt het SCE-UA-algoritme geen rekening met parameteronzekerheid. De evaluatieregel in stap (4) is zo geconstrueerd dat er alleen maar betere dan bestaande parametercombinaties worden geaccepteerd. Hoewel deze regel in essentie convergentie in de parameterruimte faciliteert, is het onrealistisch aan te nemen dat de uiteindelijk optimale parameterschatting geen fout heeft. Vanuit dit perspectief heb ik met collega's in Amsterdam en Arizona het SCE-UA-algoritme aangepast en uitgebreid, zodat het algoritme niet meer convergeert naar een enkele parameteroplossing, maar naar een posterior-verdeling (SCEM-UA: Vrugt e.a., 2003a). Deze verdeling bestaat uit een grote reeks van parametercombinaties die binnen een toegestane afstand liggen van de optimale schatting. Deze toegestane afstand wordt bepaald door Bayesiaanse theorie. De posterior-verdeling bevat belangrijke informatie over parameteronzekerheid, parametercorrelatie, en hieraan nauw gerelateerd, de informatie-inhoud van de kalibratiedata (Vrugt e.a., 2001, 2002, 2004).

Bij het optimaliseren van modelparameters wordt veelal de gehele kalibratiedataset in één doelfunctie gebruikt. Op deze wijze worden temporele variaties in optimale parameterwaarde onderdrukt, en dus de informatie-inhoud van een tijdserie gereduceerd tot een enkel getal. Een interessant alternatief is, om de kalibratiedata recursief te assimileren en de bijbehorende parameteronzekerheid als functie van toenemende lengte van de kalibratieset te bestuderen. Uit een dergelijke analyse kan worden opgemaakt welke metingen het meest informatief zijn voor de identificatie van de modelparameters. Recursieve toepassing van het SCEM-UA-algoritme op metingen van doorval en de waterberging

op vegetatie, heeft aangetoond wanneer gemeten moet worden om welke modelparameters vast te stellen, en dat doorvalmetingen slechts weinig informatie bevatten voor de kalibratie van vegetatie-interceptiemodellen (Vrugt e.a., 2003c). Verder heeft een recursieve SCEM-UA-applicatie op een gemeten tijdreeks van dagelijkse rivierafvoer geïllustreerd dat slechts 5% van de totale tijdreeks nodig is voor een goede kalibratie van een rivierafvoermodel (Vrugt e.a., 2002).

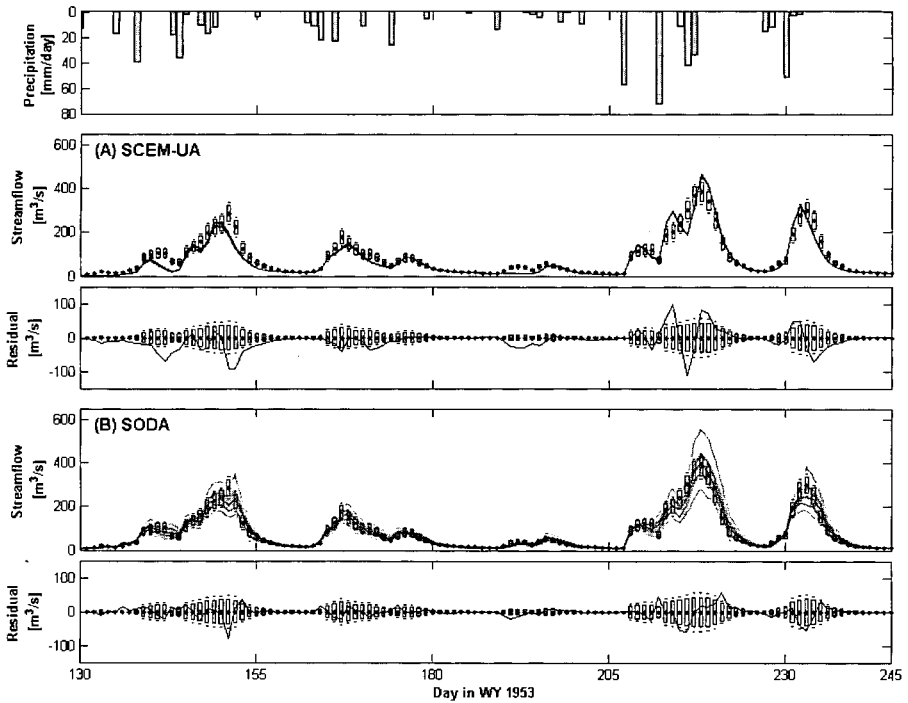
Ondanks het feit dat er in de afgelopen decennia veel vooruitgang is geboekt in de ontwikkeling en toepassing van globale optimalisatietechnieken om parameters te schatten in hydrologische modellen, hebben deze technieken duidelijke tekortkomingen. Gedurende de kalibratie van modelparameters wordt namelijk verondersteld dat de misfit tussen model en meting alleen veroorzaakt wordt door random meetfouten in de kalibratiedata. In de constructie van de doelfunctie, zoals in vergelijking (1) wordt dus impliciet aangenomen dat er geen fouten zijn in het model, de gemeten randvoorwaarden (input tot model) en de optimale parameterwaarde zelf. In de context van hydrologisch modelleren is dit natuurlijk verre van realistisch.

In het laatste hoofdstuk van mijn proefschrift heb ik daarom een gecombineerde globale optimalisatie- en data-assimilatiemethode ontwikkeld, genaamd SODA, die expliciet rekening houdt met modelfouten, input-, output- en parameterfouten tijdens de kalibratie van hydrologische modellen (Vrugt e.a., 2005a, b). De SODA-methode combineert de kracht en computationele efficiëntie van het Ensemble Kalman Filter (EnKF: Evensen, 1994) met de zoek efficiëntie van de stochastische SCEM-UA-methode om simultaan modeltoestanden te updaten en parameters te kalibreren. De SCEM-UA-methode genereert een parametercombinatie, die vervolgens door de EnKF wordt gebruikt om recursief toestandscorrectie te

doen en bijbehorende voorspellingen te genereren. Deze set van voorspellingen wordt in SCEM-UA gebruikt om de bijbehorende posterior-verdeling van de parameters te schatten.

Figuur 3 illustreert een toepassing van het SCEM-UA (3A) en SODA-algoritme (3B) op het Sacramento Soil Moisture Accounting model (SAC-SMA: 16 parameters en 9 toestanden) dat in Amerika intensief wordt gebruikt om de afvoer van rivieren te voorspellen. De boxplots geven de omvang van de meetfout van de afvoer aan, terwijl de lichte en donkere grijstone de onzekerheid in de modelvoorspelling als gevolg van model/invoerfout en parameterfout illustreren. In het geval van de SCEM-UA-methode zijn de onzekerheidsmarges erg klein, omdat modelfouten en invoerfouten verwaarloosd worden. Indien op een juiste wijze alle fouten in de kalibratie van het SAC-SMA-model worden meegenomen resulteert dit in de onzekerheidsmarges van figuur 3B. Interpretatie van de residuen (verschil tussen model en meting) als functie van tijd, illustreert dat recursieve toestandscorrecties (zoals in SODA) de autocorrelatie in de residuen significant doet afnemen.

Dat was mijn promotieonderzoek. Maar waar ben ik dan nu mee bezig in Los Alamos? Momenteel probeer ik de bestaande algoritmen uit te breiden zodat ze toepasbaar worden voor zeer hoogdimensionale problemen met een nog groter aantal modeltoestanden en parameters. Om dit mogelijk te maken, richt een deel van mijn onderzoek zich nu op het rekenen op supercomputers. Om de nodige feedback te krijgen vanuit case studies ben ik bezig met de toepassing van de ontwikkelde optimalisatie- en data-assimilatietechnieken in verschillende vakgebieden. Zo ben ik nu bezig met het voorspellen van vogeltrekroutes over Centraal-Europa met behulp van multi-objective optimalisatie (Vrugt e.a., 2003b).



Figuur 3: Met SAC-SMA gesimuleerde hydrograaf-onzekerheidsintervallen verkregen met de SCEM-UA (A), en SODA (B)-methode. De zwarte lijn is de voorspelling van de beste parameterset. De lichte en donkere grijslijnen indiceren onzekerheid in modelvoorspelling als gevolg van model- en parameterfout. De box-plots geven de omvang van de meetfout van de rivierafvoermeting weer.

Literatuur

- Duan, Q., S. Sorooshian en V.K. Gupta (1992)** Effective and efficient global optimization for conceptual rainfall-runoff models; in: *Water Resources Research*, vol 28 (4), pag 1015–1031.
- Evensen, G. (1994)** Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics; in: *J. Geophys. Res.*, vol 99 (C5), pag 10143–10162.
- Vrugt, J.A., A.H. Weerts en W. Bouten (2001)** Information content of data for identifying soil hydraulic parameters from outflow experiments; in: *Soil Sci. Soc. Am. J.*, vol 65, pag 19–27.
- Vrugt, J.A., W. Bouten, H.V. Gupta en S. Sorooshian (2002)** Toward improved identifiability of hydrologic model parameters: The information content of experimental data; in: *Water Resources Research*, vol 38 (12), pag 1312, doi:10.1029/2001WR001118.
- Vrugt, J.A., H.V. Gupta, W. Bouten en S. Sorooshian (2003a)** A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters; in: *Water Resources Research*, vol 39 (8), pag 1201, doi:10.1029/2002WR001642.
- Vrugt, J.A., H.V. Gupta, L.A. Bastidas, W. Bouten en S. Sorooshian (2003b)** Effective and efficient algorithm for multi-objective optimization of hydrologic models; in: *Water Resources Research*, vol 39 (8), pag 1214, doi:10.1029/2002WR001746.
- Vrugt, J.A., S.C. Dekker en W. Bouten (2003c)** Identification of rainfall interception model parameters from measurements of throughfall and forest canopy storage; in: *Water Resources Research*, vol 39 (9), pag 1251, doi:10.1029/2003WR002013.
- Vrugt, J.A., G.H. Schoups, J.W. Hopmans, C.H. Young, W. Wallender, T. Harter en W. Bouten (2004)** Inverse modeling of large-scale spatially distributed vadose zone properties using global optimization; in: *Water Resources Research*, vol 40(6), W06503, pag???, doi:10.1029/2003WR002706.
- Vrugt, J.A., C.G.H. Diks, W. Bouten, H.V. Gupta en J.M. Verstraten (2005a)** Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation; in: *Water Resources Research*, vol 41(1), pag ???, W01017, doi:10.1029/2004WR003059.
- Vrugt, J.A., H.V. Gupta, B. Ó Nualláin en W. Bouten (2005b, ingezonden)** Real-time data assimilation for operational ensemble streamflow forecasting; in: *J. Hydrometeor.*

Jasper A. Vrugt

Director's funded Postdoctoral Fellow
Earth and Environmental Sciences

Division

Los Alamos National Laboratory

E-mail: vrugt@lanl.gov