Arnold Heemink
Faculteit EWI
TU Delft
a.w.heemink@tudelft.nl

Peter Jan van Leeuwen
Department of Meteorology
University of Reading, UK
p.j.vanleeuwen@reading.ac.uk

Voorspellen met behulp van data-assimilatie

Voor het maken van voorspellingen van bijvoorbeeld het weer zijn behalve een numeriek model ook waarnemingen van groot belang. Onder meteorologen staat het systematisch combineren van modellen en waarnemingen bekend als data-assimilatie. Moderne technieken voor data-assimilatie maken gebruik van ideeën uit de optimale besturingstheorie en de filtertheorie, en het ontwikkelen van dergelijke technieken is een actief onderzoeksterrein waar zowel wiskundigen als meteorologen en oceanografen aan werken. Arnold Heemink en Peter Jan van Leeuwen bespreken in dit artikel de voornaamste hedendaagse data-assimilatie methoden.

Numeriek wiskundige modellen worden tegenwoordig steeds vaker gebruikt bij het berekenen van voorspellingen. Voorbeelden hiervan zijn voorspellingen van het weer, de waterstanden langs de Nederlandse kust of luchtvervuiling in steden. De resultaten van de modelberekeningen zijn echter beïnvloed door tal van onzekerheden. Bijvoorbeeld door onnauwkeurige begin- en of randvoorwaarden of door onzekerheden in de waarden van de modelparameters. Door gebruik te maken van waarnemingen kunnen de modelresultaten worden verbeterd. Dit wordt data-assimilatie genoemd. Deze term is vele jaren geleden al geïntroduceerd door de meteorologen. Meteorologen zeggen vaak dat zij het weer van morgen niet goed kunnen voorspellen om de simpele reden dat het weer van vandaag niet overal goed bekend is. Veel fouten in de kortetermijnmodelvoorspellingen zijn het gevolg van fouten die op het moment van voorspellen al aanwezig waren. Dit geeft het belang aan van goede meetinformatie én van goede data-assimilatietechnieken om de voorspelfouten te minimaliseren.

Tot het eind van de negentiger jaren was de meest gebruikte data-assimilatietechniek optimale interpolatie. Hierbij werd de berekende modeltoestand gecorrigeerd op basis van waarnemingen van die toestand en aannames omtrent de statistiek van de modelfout en de meetfout. Optimale interpolatie heeft een zeer principieel nadeel. De gecorrigeerde modeltoestand is niet consistent met het model omdat de procesdynamica niet gebruikt wordt bij het bepalen van de correcties. De gecorrigeerde modeltoestand voldoet dan niet aan de modelvergelijkingen, wat zelfs kan leiden tot instabiliteiten.

Data-assimilatie kan gezien worden als een invers probleem. Op basis van de uitvoer van het model op de meetlocaties worden de modelfouten gereconstrueerd. Met informatie over het gevolg moet dus de oorzaak worden opgespoord. De algoritmen die voor nauwkeurige data-assimilatieschema's gebruikt worden, zijn sterk geïnspireerd door de optimale besturingstheorie en de filtertheorie. De laatste aanpak resulteert uiteindelijk in het Ensemble Kalman Filter. Anders dan optimale interpolatie zijn deze methoden in staat om het inverse probleem netjes op te lossen, waardoor de modeltoestand ook na correctie blijft voldoen aan de modelvergelijkingen. Het gebruik van deze data-assimilatiemethoden is daardoor niet slechts een marginale verbetering ten opzichte van een traditionele aanpak als optimale interpolatie. Het is een wezenlijke stap vooruit.

Data-assimilatie kan ook worden geformuleerd op basis van het Theorema van Bayes. Voor het geval dat modelonzekerheden benaderd kunnen worden met behulp van een Gaussische kansdichtheid resulteert dit ook weer in het Ensemble Kalman Filter. Als de Gaussische benadering niet op gaat, krijgen we het Particle Filter.

4D variationele data-assimilatie

Beschouw het volgende numeriek wiskundige model:

$$\begin{aligned} x(t_{i+1}) &= M_i[x(t_i), p], \\ x(t_0) &= x_0(p), \quad i = 0, 1, 2, \ldots, K, \end{aligned} \tag{1}$$

waarbij $x(t_i)$ de toestandsvector is op tijdstip t_i en p de te schatten parametervector. De niet-lineaire functie M_i stelt het numerieke schema voor. $M_i[x(t_i), p]$ is een representatie van een computercode die op een ingewikkelde wijze $x(t_{i+1})$ uitrekent bij gegeven input $x(t_i)$ en p. $x(t_i)$ bestaat uit alle variabelen in alle roosterpunten van het model. De dimensie van $x(t_i)$ kan daarom zeer groot zijn $(10^6 - 10^8)$. De parameters kunnen ruimteafhankelijke fysische parameters zijn in het model, parameters in de randvoorwaarden, maar ze kunnen ook de volledige begincondition bevatten. Ook de dimensie van p kan dus heel groot zijn.

Er wordt vervolgens aangenomen dat de waarnemingen beschikbaar zijn in de volgen-

Minimalisatie met het geadjungeerde model

Het principe van de geadjungeerde aanpak is eenvoudig. Beschouw bijvoorbeeld het volgende niet-lineaire model:

$$x(t_{i+1}) = M_i[x(t_i)],\ x(t_0) = p\ ,\quad i = 0, 1, \ldots, K, \tag{2}$$

en de situatie dat er meetinformatie beschikbaar is van $x(t_K)$. We willen nu de beginvoorwaarde p schatten via het minimaliseren van de doelfunctie

$$J(x(t_K)) = J(M_{K-1}[M_{K-2}[M_{K-3}[\cdots[M_0[p]]\cdots]]]). \tag{3}$$

We kunnen nu eenvoudig het effect δJ uitrekenen van een kleine verstoring δp in de parameter op de waarde van de doelfunctie:

$$\begin{aligned}\delta J &= \left[\frac{\partial J}{\partial x(t_K)}\right]^T \cdot \frac{\partial M_{K-1}}{\partial x(t_{K-1})} \cdot \cdots \cdot \frac{\partial M_0}{\partial x(t_0)} \cdot \delta p \\ &= \left[\frac{\partial J}{\partial x(t_K)}\right]^T \cdot M(t_K, t_{K-1}) \cdot \cdots \cdot M(t_1, t_0) \cdot \delta p,\end{aligned} \tag{4}$$

met $M(t_i, t_{i-1})$ het gelineariseerde model.
We kunnen dit herschrijven als een inwendig product:

$$\delta J = \left\langle \frac{\partial J}{\partial x(t_K)}, M(t_K, t_{K-1}) \cdots M(t, t_0)\delta p \right\rangle . \tag{5}$$

Bij veel parameters is dit geen efficiënte wijze om het effect van een verstoring δp op de doelfunctie te berekenen. Bij iedere δp moet een volledige simulatie van het gelineariseerde model worden uitgevoerd.
Maar we kunnen vergelijking (5) ook schrijven als

$$\delta J = \left\langle M^*(t, t_0) \cdots M^*(t_K, t_{K-1}) \cdot \frac{\partial J}{\partial x(t_K)}, \delta p \right\rangle \equiv \left\langle \frac{\partial J}{\partial p}, \delta p \right\rangle , \tag{6}$$

waarbij $M^*(t_i, t_{i-1})$ de geadjungeerde operator is van $M(t_i, t_{i-1})$. Aangezien $M(t_i, t_{i-1})$ een matrix is, is $M^*(t_i, t_{i-1})$ niets anders dan de getransponeerde van deze matrix. Vergelijking (6) is een zeer efficiënte uitdrukking voor de gradiënt van de doelfunctie. Met slechts één berekening van het geadjungeerde model terugwaarts in de tijd kan deze gradiënt berekend worden, ongeacht het aantal parameters.

de vorm:

$$y_i^0 = H_i[x(t_i)] + \epsilon_i,\quad i = 1, 2, \ldots, K, \tag{7}$$

waarbij y_i^0 de waarnemingen zijn op tijdstip t_i. De functie H_i is de relatie tussen de metingen en de toestandsvector.

De onbekende parameters kunnen geschat worden door de volgende doelfunctie te minimaliseren:

$$\begin{aligned}J_{(p)} =& \frac{1}{2}(p - p_0)^T P_0^{-1}(p - p_0) \\ &+ \frac{1}{2}\sum_{i=1}^{K}(H_i[x(t_i)] - y_i^0)^T \\ &\cdot R_i^{-1}(H_i[x(t_i)] - y_i^0).\end{aligned} \tag{8}$$

Hierbij is de eerste term een regularisatieterm, terwijl de tweede uitdrukking de afstand van het model tot de beschikbare waarnemingen representeert. De modelvergelijking (1) is bij het minimaliseren van (8) een nevenvoorwaarde.

Het bovenstaande probleem is een standaard optimalisatieprobleem. Via de introductie van Lagrange-multiplicatoren of geadjungeerde toestanden kunnen de nevenvoorwaarden aan de doelfunctie worden toegevoegd. Vervolgens kan via variatierekening een eenvoudige uitdrukking van de gradiënt van de doelfunctie worden verkregen. Hierbij is dan wel de oplossing nodig van het geadjungeerde model van het gelineariseerde model (zie het kader).

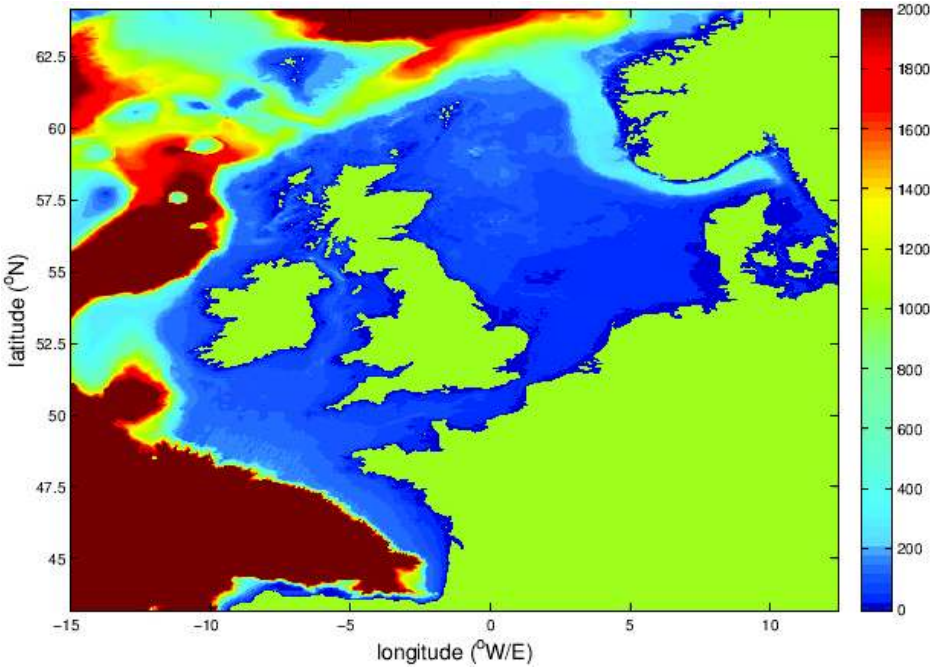
Door de bovenstaande aanpak te combineren met een gradiëntmethode voor het minimaliseren van de doelfunctie, kan er een efficiënte data-assimilatiemethode worden verkregen. Deze methode wordt 4D-Var genoemd. Een groot praktisch nadeel van variationele data-assimilatie is dat het geadjungeerde model beschikbaar moet zijn. Dit is bij zeer complexe modelsystemen een grote programmeerinspanning, vaak enkele mensjaren programmeerwerk. Er zijn geadjungeerde compilers beschikbaar, die als invoer de code van een voorwaarts model vragen en dan als output de code van het gelineariseerde model of van het geadjungeerde model produceren. Deze compilers werken echter alleen goed bij relatief eenvoudige modellen.

Een fundamenteel probleem is dat het optimalisatieprobleem niet convex is en er dus een risico is dat het algoritme strandt in een lokaal minimum. Voorts is het algoritme zeer rekenintensief en kunnen we slechts een beperkt aantal iteraties uitvoeren en zullen dan tevreden moeten zijn met het resultaat. Ondanks deze fundamentele problemen heeft 4D-Var zijn grote waarde in de praktijk bewezen en wordt het op grote schaal met succes toegepast.

Een alternatief voor de standaard 4D-Var aanpak is *incremental 4D-Var*. Hierbij wordt een vereenvoudigd model gebruikt voor de assimilatie. Dit vereenvoudigde model wordt gelineariseerd met als referentie de beste voorspelling van het oorspronkelijke model. Vervolgens wordt de minimalisatie uitgevoerd met dit vereenvoudigde en gelineariseerde model. Het uiteindelijke resultaat levert een nieuwe voorspelling op. Deze is echter niet optimaal omdat gebruik is gemaakt van het vereenvoudigde model. Door de procedure nu opnieuw uit te voeren met de nieuwe voorspelling als referentie, kan het resultaat nog verbeterd worden. Convergentie is hierbij helaas niet gegarandeerd. Incremental 4D-Var is de meest gebruikte methode voor het berekenen van de dagelijkse weersvoorspellingen.

Zeer recent is de aandacht verschoven naar een aanpak analoog aan incremental 4D-Var, waarbij in plaats van het ontwikkelen van een nieuw vereenvoudigd numeriek model gebruik wordt gemaakt van *Proper Orthogonal Decomposition* om een vereenvoudigd model te verkrijgen. Met deze generieke modelreductie-aanpak hoeft er dus geen apart vereenvoudigd numeriek model ontwikkeld te worden. Deze aanpak is recent bijvoorbeeld gebruikt bij het schatten van de bodemtopografie van het nieuwe stormvloedvoorspelmodel van Deltares.

De bodemgegevens van een stormvloedmodel worden meestal verzameld ten behoeve van de scheepvaart en de bodemkaarten die daaruit worden afgeleid, hebben de nei-



Figuur 1 De bodemtopografie van het domein van het stormvloedvoorspelmodel

ging wat te ondiep te zijn. Als deze kaarten vervolgens worden gebruikt voor het bepalen van de geometrie van het numerieke stormvloedmodel zullen de getijgolven in dit model zich te langzaam voortplanten. Door gebruik te maken van de vele meetgegevens die van het getij beschikbaar zijn, kan de bodemtopografie in het model gecorrigeerd worden net zolang totdat de golfvoortplanting in het model correct is. In Figuur 1 is het domein van het nieuwe stormvloedmodel en de bodemtopografie voor calibratie weergegeven. In Figuur 2 zijn de *root-mean-square errors* in een aantal meetstations voor en na calibratie weergegeven. Bij dit calibratie-experiment zijn twintig correctieparameters geschat. De rekentijd die nodig was om dit inverse probleem op te lossen was equivalent met slechts elf evaluaties van de doelfunctie (=model simulaties). Dus behalve dat het geadjungeerde model niet nodig is, is de aanpak door deze toepassing ook zeer efficiënt gebleken.

Ensemble data-assimilatie

Data-assimilatie kan ook beschreven worden in termen van waarschijnlijkheidsdichtheden, of 'probability density functions', kortweg pdf's. Dit werkt als volgt: de onzekerheid in de modelvoorspellingen geven we weer met een pdf $p(x)$. Deze pdf is het resultaat van onzekerheden in de begintoestand van de voorspelling, in de modelvergelijkingen (die numerieke discretisaties van niet volledig bekende fysische wetten zijn), en in de randvoorwaarden. Data-assimilatie wordt nu gezien als het aanpassen van deze 'a-priori pdf' voor de waarnemingen, resulterende in de 'a-posteriori pdf', de pdf van de mogelijke modeltoestanden gegeven die waarnemingen $p(x|y)$. In wiskundige notatie vinden we

$$p(x|y) = \frac{p(y|x)}{p(y)} p(x). \qquad (9)$$

Deze relatie, het Theorema van Bayes, vormt de algemene basis voor data-assimilatie. Het geeft weer hoe onze kennis van het systeem, gegeven via de a-priori pdf $p(x)$, wordt gemodificeerd door de waarnemingen y, tot de a-posteriori pdf $p(x|y)$. De prior wordt vermenigvuldigd met de zogenaamde likelihood $p(y|x)$, die voor iedere modeltoestand x aangeeft hoe waarschijnlijk de waarnemingen y zijn.

Ten eerste valt op dat data-assimilatie nu te interpreteren is als een vermenigvuldigingsprobleem: we moeten de a-priori pdf vermenigvuldigen met de zogenaamde likelihood $p(y|x)$ om de a-posteriori pdf te vinden. De noemer $p(y)$ kan gezien worden als een normalisatiefactor omdat de waarnemingen al bekend zijn.

Ondanks het belang van het hebben van een formele oplossing voor het data-assimilatieprobleem, is de praktijk weerbarstig. Het probleem is dat de modellen die gebruikt worden in de geowetenschappen typisch zeer hoge dimensies hebben, en alleen al voor het opslaan van de a-priori pdf hebben we computers nodig met astronomisch grote geheugens. We zijn dus gedwongen benaderingen te aanvaarden. Een veelgebruikte benadering is dat de pdf Gaussisch is, zodat we alleen de gemiddelde modeltoestand en de bijbehorende covariantie nodig hebben. Dit leidt onmiddellijk naar het zogenaamde Kalman Filter. Helaas is het voor de hoogdimensionale systemen niet mogelijk de covariantie op te slaan, dus we moeten extra aanpassingen toestaan. Een zeer populaire is het Ensemble Kalman Filter. Bij deze methode wordt de pdf weergegeven als een ensemble van modeltoestanden. Iedere modeltoestand evolueert volgens de modelvergelijkingen naar de tijd waarop we waarnemingen hebben. Daar berekenen we het gemiddelde van de Gaussische pdf als het gemiddelde van de ensembleleden, en ook de covariantie wordt berekend gebruikmakende van het ensemble van modeltoestanden.

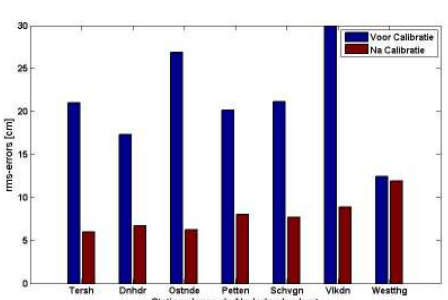
Deze methode is erg succesvol en wordt tegenwoordig gebruikt in zeer hoogdimensionale systemen. Figuur 3 geeft een voorbeeld van het gebruik van het Ensemble Kalman Filter in een model van de Atlantische Oceaan. Het model heeft meer dan een miljoen variabelen op ieder tijdstip, en we assimileren satellietwaarnemingen van de temperatuur van het zeewater en de hoogte van het zeeoppervlak. Die laatste waarnemingen geven informatie over het drukveld in de oceaan, welke zeer belangrijk is in het beïnvloeden van stromingspatronen.

Een nieuwe ontwikkeling is het loslaten van de restrictie tot Gaussische pdf's. De methode, Particle Filter genoemd, gebruikt weer een ensemble van modeltoestanden en gebruikt het Theorema van Bayes meer direct. In wiskundige termen, de a-priori pdf wordt weergegeven als een ensemble van modeltoestanden x_i, als

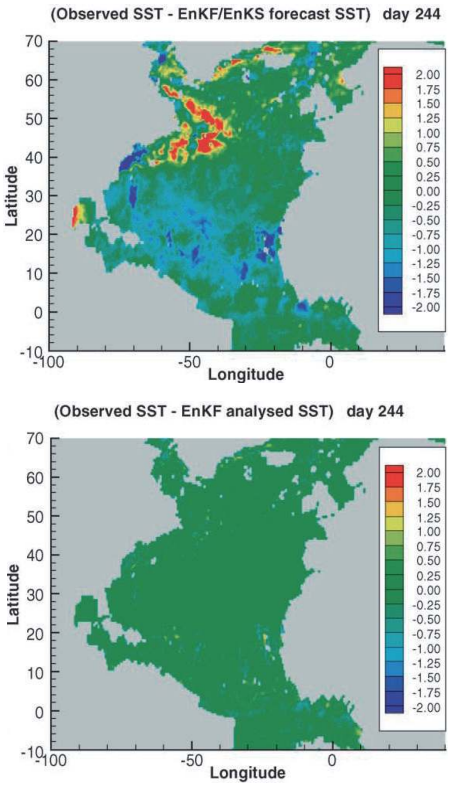
$$p(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - x_i). \qquad (10)$$

Als we dit direct gebruiken in het Theorema van Bayes vinden we

$$p(x|y) = \sum_{i=1}^{N} w_i \delta(x - x_i), \qquad (11)$$



Figuur 2 De RMS error van de modelresultaten voor en na calibratie

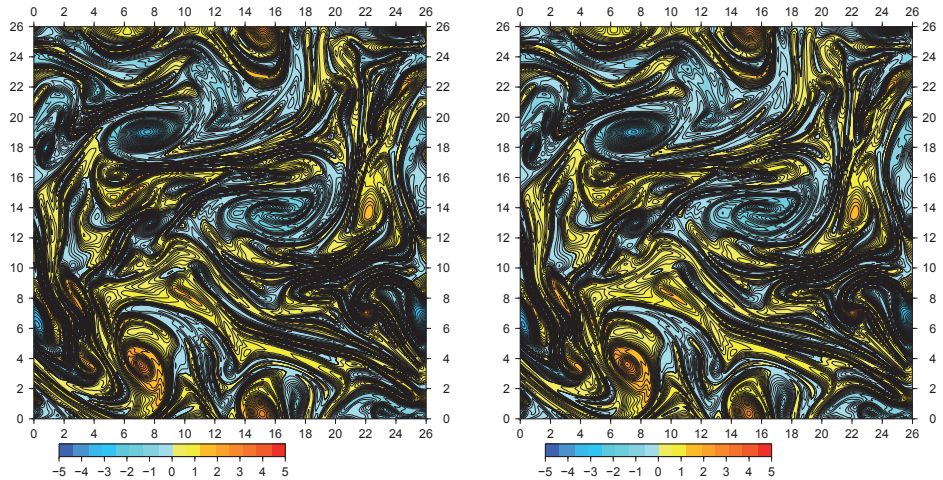


Figuur 3 Voorbeeld van succesvolle assimilatie in de Atlantische Oceaan. Het bovenste plaatje geeft het verschil tussen een 7-daagse voorspelling voor zeewatertemperatuur (SST) van het model en de waarnemingen weer op dag 244. Verschillen van 2 graden Celsius zijn zichtbaar, bijvoorbeeld voor de Amerikaanse kust. Het onderste plaatje geeft het verschil tussen de waarnemingen en het geassimileerde model weer op dag 244. De verschillen zijn minimaal, wat aangeeft dat bijna alle informatie in de waarnemingen opgenomen is in het model.

waarbij de gewichten w_i gegeven zijn als

$$w_i = \frac{p(y|x_i)}{\sum_j p(y|x_j)}, \qquad (12)$$

wat laat zien dat de waarde van de gewichten w_i afhangt van hoe waarschijnlijk modeltoestand x_i is als de waarnemingen gegeven zijn als y. Hoe dichter modeltoestand x_i bij de waarnemingen is, hoe groter zijn gewicht. De formule voor de a-posteriori pdf laat zien dat de oplossing van het data-assimilatieprobleem een gewogen ensemble is. Bijvoorbeeld voor het vinden van de gemiddelde toestand nemen we het gewogen



Figuur 4 Snapshot van het vorticiteitsveld van de 'waarheid' (links) en de schatting van het Particle Filter (rechts). Merk op dat de verschillen minimaal zijn, wat aangeeft dat het Particle Filter zeer goed werkt.

gemiddelde over de ensembleleden, met gewichten w_i.

Het Particle Filter als hierboven weergegeven is niet erg effectief voor systemen met dimensies groter dan 10. Het probleem is dat de gewichten te veel uit elkaar liggen, met typisch een van de gewichten dicht bij 1, en al de andere gewichten zeer dicht bij nul. Het gewogen ensemble bestaat dan effectief uit een modeltoestand, en alle informatie over de pdf is verloren. Gelukkig is de ontwikkeling van efficiëntere Particle Filters in volle gang. Er bestaan nu methoden die efficiënt zijn voor systemen met zeer grote dimensie. Een voorbeeld is het zogenaamde Equivalent-Weight Particle Filter. In dit filter worden de modeltoestanden zo aangepast dat ze bijna gelijke gewichten hebben, terwijl we toch het data-assimilatieprobleem netjes oplossen.

Figuur 4 geeft een voorbeeld van het toepassen van het Equivalent-Weights Particle Filter in een turbulent sterk niet-lineair systeem van cyclonale en anticyclonale wervels zoals die voorkomen in de oceaan. Dit voorbeeld is van een test van de methode waarbij de waarnemingen verkregen zijn van een modelrun met verstoorde begincondities en realisaties van modelfouten. De dimensie van het systeem is 65.000, en in dit voorbeeld is het vorticiteitsveld waargenomen op ieder roosterpunt iedere 50 modeltijdstappen. De decorrelatietijd van het vorticiteitsveld is 25 modeltijdstappen, wat aangeeft dat dit een moeilijk en sterk niet-lineair data-assimilatieprobleem is. Figuur 4 laat zien dat de waarheid en de schatting ven het gemiddelde van de ensembleleden van het Particle Filter zeer dicht bij elkaar liggen. Naast andere statistische technieken geeft dit vertrouwen in de data-assimilatiemethode.

Tot slot

Numerieke modellen worden steeds beter en steeds vaker gebruikt voor praktische toepassingen. Er wordt ook steeds meer meetinformatie ingewonnen. Het gevolg is dat het gebruik van data-assimilatietechnieken de afgelopen tijd zeer sterk is toegenomen. Deze ontwikkeling is nog verder versneld doordat er inmiddels enkele goede ensembletechnieken beschikbaar zijn die vrij eenvoudig kunnen worden gekoppeld aan bestaande numerieke modelsystemen. Van veel ensemblealgoritmen is ook open source code te downloaden (zie www.openda.org) zodat het benodigde programmeerwerk heel beperkt is. Daarnaast zijn er ook nog vele wetenschappelijke uitdagingen om de methoden efficiënter en nauwkeuriger te maken, met name voor sterk niet-lineaire problemen.

Referenties

1 M.U. Altaf, M. Verlaan en A.W. Heemink, Efficient identification of uncertain parameters in a large-scale tidal model of the European continental shelf by proper orthogonal decomposition, *Int. J. Numer. Meth. Fluids* 68 (2012), 422–450.

2 K. Brusdal, J.M. Brankart, G. Halberstadt, G. Evensen, P. Brasseur, P.J. van Leeuwen, E. Dombrowsky en J. Verron, A demonstration of ensemble-based assimilation methods with a layered OGCM from the perspective of operational ocean forecasting systems, *Journal of Marine Systems* 40–41 (2003), 253–289.

3 P. Courtier, J.N. Thepaut en A. Hollingsworth, A strategy for operational implementation of 4D-VAR, using an incremental approach, *Quarterly Journal of the Royal Meteorological Society* 120 (1994), 1367–1387.

4 G. Evensen, *Data Assimilation: The Ensemble Kalman Filter*, Second Edition, Springer, 2009.

5 F.X. Le Dimet en O. Talagrand, Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects, *Tellus* 38 (1986), 97–110.

6 P.J. van Leeuwen, Particle Filtering in Geophysical Systems, *Monthly Weather Review* 137 (2009), 4089–4114.

7 P.J. van Leeuwen en M. Ades, Efficient fully non-linear data assimilation for geophysical fluid dynamics, *Computers and Geosciences* 55 (2013), 16–27.