

Kapitel 15 VALIDERING

Jens Christian Refsgaard
Hydrologisk afdeling, GEUS

Nøglebegreber: Validering af model, verifikation af kode, valideringstests, split-sample, proxy-basin, differential split-sample, overparameterisering og frihedsgrader, uafhængige datasæt, jackknifing og modelgyldighed.

ABSTRACT: Spørgsmålet om, hvorvidt modeller overhovedet kan valideres har været omdiskuteret det seneste årti, og udfra et anti-positivistisk synspunkt, kan en teori (her en model) aldrig bevises som værende generel gyldig, men kan derimod afvises hvis et enkelt eksempel begrundet dette. Et vigtigt element i modelvalideringen er fastsættelsen af relevante test. For at vurdere om en kalibreret model kan betragtes som gyldig til efterfølgende brug skal den valideres ved test mod andre data end dem, der blev brugt til kalibrering. Model validering indebærer således at dokumentere, at en sted-specifik model kan producere beregningsresultater, som kan opfylde de givne nøjagtighedskrav. En model må kun karakteriseres som valideret på de områder, hvor der er gennemført valideringstest.

15.1 INDLEDNING

Begreberne validering og verifikation bliver brugt i flæng (men i forskellig betydning) hos forskellige forfattere, f.eks. Konikow (1978), Schlesinger et al. (1979), Klemes (1986), Tsang (1991), Anderson and Woessner (1992), Flavelle (1992), Larsson (1992), Oreskes et al. (1994), Dee (1995), Refsgaard and Knudsen (1996) og Beven (1996).

Spørgsmålet om, hvorvidt modeller overhovedet kan valideres er med stærke bånd til grundlæggende videnskabsfilosofiske opfattelser blevet diskuteret voldsomt de senere år. Konikow og Bredehoeft (1992) fremførte således som eksponent for den ene opfattelse at begreberne validering og verifikation er vildledende og slet ikke burde bruges:

“ .. the terms validation and verification have little or no place in ground-water science; these terms lead to a false impression of model capability”

Hovedargumentationen herfor er det anti-positivistiske synspunkt, at en teori (her en model) aldrig kan bevises som værende generel gyldig, men derimod kan afkræftes ved et enkelt eksempel. De Marsily et al. (1992) argumenterer i et svar til Konikow og Bredehoeft (1992) for et mere pragmatisk synspunkt:

“ .. using the model in a predictive mode and comparing it with new data is not a futile exercise; it makes a lot of sense to us. It does not prove that the model will be correct for all circumstances, it only increases our confidence in its value. We do not want certainty; we will be satisfied with engineering confidence.”

Et andet eksempel på en forsker, som er kritisk overfor begrebet modelvalidering er Beven (1996), der bl.a. som en reaktion på at mange modeludviklere og -brugere i tidens løb har benyttet ‘validering’ næsten som et gummistempel, påstod at modelvalidering ikke er mulig:

“ .. is there any example of a successfully validated distributed model at the catchment scale ?”

Et hovedproblem i denne diskussion er, at begrebet model ofte benyttes i to vidt forskellige betydninger, nemlig modelkode og en model opstillet for et givet område. Konikow and Bredehoeft (1992), Beven (1996) og andre modstandere mod at bruge ordet modelvalidering referer til begre-

bet modelkode, mens Schlesinger et al. (1979), Anderson and Woessner (1992) og andre forfattere, som benytter begrebet modelvalidering, refererer til model i den mindre generelle betydning en sted-specifik model.

Det vil formentlig aldrig være muligt, og måske heller ikke ønskværdigt, at opnå enighed i forskerkredse om dette grundlæggende videnskabsteoretiske spørgsmål. Dette giver selvfølgelig nogle grundlæggende problemer med hensyn til at nå til enighed i den praktiske verden om en fælles anderkendt terminologi og metodik for modelvalidering. Den praktiske brug af modeller vil øges meget fremover, og mange brugere benytter begrebet modelvalidering uden måske at have en præcis definitionen. Så der vil under alle omstændigheder være brug for at opnå en eller anden form for stramning af terminologi og metodik samt en øget bevidsthed om nødvendigheden heraf.

15.2 DEFINITION OG FORMÅL MED MODELVALIDERING

Som beskrevet i det indledende kapitel (tabel 1.1) benyttes følgende begreber i nærværende ståbi:

- *Verifikation* benyttes om en modelkode, dvs. dokumentation for at en modelkode løser de givne ligninger med en specificeret nøjagtighed.
- *Validering* benyttes i relation til en sted-specifik model, dvs. dokumentation for at en sted-specifik model kan beskrive virkeligheden med en tilfredsstillende nøjagtighed.

En model bør aldrig anvendes uden en forudgående modelvalidering, fordi modellens pålidelighed i så fald ikke kendes. Formålet med modelvalidering er derfor, ved gennemførelse af passende test, at undersøge og dokumentere hvilken nøjagtighed en given model kan forventes at have, når den efterfølgende skal anvendes.

Et vigtigt element i modelvalideringen er fastsættelsen af relevante test. I næste afsnit er der foreslået nogle generelle testtyper, hvor det gennemgående princip er, at et valideringstest skal indeholde situationer sammenlignelige med dem modellen bliver udsat for i den efterfølgende model-anvendelse.

15.3 FORSKELLIGE TYPER AF VALIDERINGSTEST

Grundvandsmodeller indeholder et meget stort antal parametre, hvorfor det næsten altid vil være muligt at finde en kombination af parameterværdier, som giver en god overensstemmelse mellem modeloutput og feltdata i en kort kalibreringsperiode. Det giver imidlertid ingen garanti for, at modelstrukturen eller de skønnede parameterværdier er passende og vil resultere i pålidelige forudsigelser for andre perioder.

For at vurdere hvorvidt en kalibreret model kan betragtes som gyldig til efterfølgende brug skal den valideres ved test mod andre data end dem, der blev brugt til kalibrering. Model validering indebærer således at dokumentere, at en sted-specifik model kan producere beregningsresultater, som kan opfylde de givne nøjagtighedskrav. Det er derfor nødvendigt at specificere sådanne nøjagtighedskrav, før kalibreringen og validerings testene gennemføres. I fastsættelsen af det acceptable niveau for nøjagtighed skal der laves en afvejning mellem hvilke yderligere omkostninger, i form af data indsamling og modelleringsarbejde, og hvilke benefits, i form af større model nøjagtighed, som større nøjagtighedskrav medfører. Nøjagtighedskriterierne vil derfor variere fra sag til sag, og bør ikke fastlægges af modelbrugeren, men af vandressourceforvalteren.

Proceduren foreslået nedenfor er baseret på Klemes (1986). De fire testtyper svarer til forskellige situationer med hensyn til, hvorvidt der findes data til at foretage en kalibrering, og hvorvidt situationen i modelområdet er stationær. Stationaritet hentyder i denne sammenhæng eksempelvis til, hvorvidt der er tale om menneskeskabte ændringer (oppumpning, ændret arealanvendelse, klimaændring, mv.) eller ikke.

- *Split-sample test* er den klassiske test, hvor de tilgængelige data mængde deles i to, hvorefter der kalibreres på den ene og efterfølgende valideres på den anden. Den kan bruges i situationer, hvor der er tilstrækkelige data til en meningsfuld kalibrering, og hvor situationen er stationær.

- *Proxy-basin testet* skal anvendes hvis der ikke findes data til kalibrering i det pågældende område Z, men situationen er stationær. I sådanne tilfælde bør der foretages test i to lignende områder, X og Y, hvor ukalibrerede (blindsimulerede) modelresultater sammenlignes med felt-data. Det er vigtigt, at proceduren for parameterestimationen i de to oplande laves objektiv og gennemskuelig, og at en lignende procedure anvendes i område Z.

Eksempel: I område Z findes data til kalibrering og split-sample test af strømningsmodellen; men der findes ingen koncentrationsdata, så det er ikke muligt at kalibrere og validere stoftransportmodulet på lokale data. Testet består så i at lave blandsimulering af stoftransport (strømningsmodellen kan kalibreres) på to andre oplande X og Y (f.eks. naboområder) med skøn af porøsitet og dispersiviteter taget fra standardtabeller eller på anden objektiv vis relateret til geologien i områderne. Såfremt testene på X og Y kommer inden for nøjagtighedskravene, kan modellen siges at være valideret på område Z.

- *Differential split-sample test*. Denne test skal anvendes i ikke-stationære situationer, hvor der findes lokale data i opland Z til kalibrering, men hvor modellen skal bruges til at forudsige konsekvenser af menneskelige indgreb som f.eks. oppumpninger, ændringer i arealanvendelse eller klimaændringer. Tilsvarende til proxy-basin testet skal der her foretages test på sammenlignelige situationer i to tilsvarende områder, X og Y.

Eksempel: I område Z skal der simuleres konsekvenser af en fremtidig grundvandsoppumpning på vandløb og trykniveauer, og der findes ikke historiske data i området med ændringer i oppumpningen af den pågældende størrelsesorden. Testet består så i at lave blandsimulering på en sammenlignelig situation i to andre oplande, X og Y, med nogenlunde tilsvarende hydrogeologi. Modellerne kan på X og Y kalibreres på situationen før indvindingen, og skal herefter simulere situationen med indvinding. Såfremt testene på X og Y kommer inden for nøjagtighedskravene, kan modellen siges at være valideret på område Z.

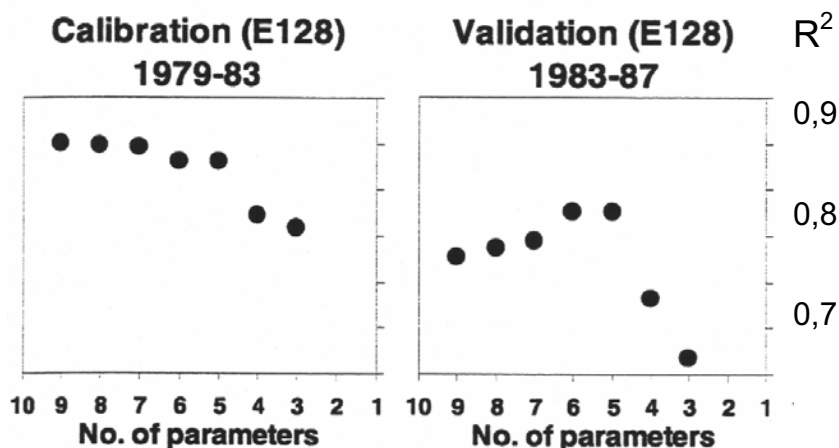
- *Proxy-basin, differential split-sample test*. Dette er den vanskeligste test for en hydrologisk model, fordi der ikke findes kalibreringsdata, og modellen skal benyttes til at vurdere en ændret fremtidig situation. Testet er en kombination af de to foregående. I praksis vil det formentlig være vanskeligt for en model at passere dette test positivt uden at benytte meget brede nøjagtighedsrammer, dvs. der vil være meget stor usikkerhed på modelforudsigelserne.

Eksempler på anvendelse af denne testprocedure til modelvalidering findes i Refsgaard and Knudsen (1996). Et andet eksempel på at definere en lignende stringent testprocedure for validering af pesticidmodeller til registreringsformål er beskrevet i Styczen (1995).

15.4 PRAKTISKE ASPEKTER I FORBINDELSE MED MODELVALIDERING

15.4.1 Overparameterisering – antal "frie" parametre

Som beskrevet i afsnit 4.6 er det meget afgørende for en models prediktionsevne at antallet af parameterværdier, som fastsættes "frit" i kalibreringen, er så lavt som muligt. Problemstillingen kan illustreres i figur 15.1.



Figur 15.1 Eksempel på en models nøjagtighed i henholdsvis kalibrerings og valideringsperiode ved valg af forskelligt antal frie parameterværdier, som tilpasses under kalibreringen (Lidén, 2000)

Figur 15.1 er fremkommet som resultat af en split-sample test, hvor der i kalibreringsperioden er foretaget automatisk kalibrering på et forskelligt antal (varierende fra 3 til 9) parametre. Figuren illustrerer tydeligt, at modeltilpasningen (jo højere R^2 værdi jo bedre model) bliver bedre jo flere frie parametre, der indrages i kalibreringen, men den viser samtidigt, at hvis parameterantallet bliver for stort, uddar kalibreringen til ren kurvefitning med ringe prediktionsevne mod uafhængige (validerings)data. Eksemplet i figur 15.1 stammer ikke fra grundvandsmodellering, men den principielle problemstilling er ligeså aktuel i grundvandsmodellering.

Konklusionen på ovenstående er ikke, at modellen ikke må have mange forskellige parameter-værdier. Der må godt være et stort antal, hvis blot størstedelen af dem fastlægges på objektiv vis på baggrund af feltdata, som er uafhængige af kalibrerings (og validerings) dataene. Når man så kommer til kalibreringen, er det vigtigt kun at indrage et begrænset antal parametre som "frie". Præcist hvor mange frie parametre, der bør indrages, kan være vanskeligt at sige, fordi det afhænger af informationsindholdet i kalibreringsdataene. Jo bedre datagrundlag i kalibreringen, jo flere parameterværdier kan man tillade sig at kalibrere på.

15.4.2 Valideringstest ved stationær grundvandsmodel

Ved en stationær grundvandsmodel foretages der kalibrering på et enkelt trykniveaubillede, som antages at være repræsentativt for en stationær situation. Derfor kan man ikke her gennemføre en traditionel split-sample test, som beskrevet i afsnit 15.2. Men det er dog stadig af afgørende betydning at gennemføre valideringstest mod uafhængige data. Der er forskellige muligheder herfor, fx.:

- Trykniveauobservationerne, som danner grundlaget for det stationære observerede grundvandsbillede, opdeles i to lige store dele. Herefter foretages der kalibrering mod den ene halvdel af dataene, mens den anden halvdel gemmes som uafhængige data i en efterfølgende valideringstest. Denne proces kan herefter gentages i omvendt rækkefølge. Herved kan opnås en vurdering af modellens evne til at simulere trykniveauer.
- Såfremt der anvendes automatisk kalibrering (invers modellering), kan man på mere systematisk vis undersøge prediktionsevnen, fx ved skiftevis at kalibrere mod alle datapunkter bortset fra et enkelt og herefter validere mod det uafhængige datapunkt. Denne testmetode benævnes jackknifing.

15.4.3 *Begrænsninger i valideret models gyldighedsområde*

Traditionelt er en model blevet testet med en split-sample valideringstest, hvorefter modellen med henvisning til dens fysisk baserede opbygning ofte er blevet erklæret generelt gyldig, også til simulering af variable og under forhold, hvorpå den slet ikke er testet. En række eksempler fra den internationale litteratur (fx Ambroise et al, 1995; Refsgaard et al., 1997; og Jayatilaka et al., 1998) viser imidlertid, at en model der eksempelvis er kalibreret mod afstrømning på en lokalitet ikke nødvendigvis er lige så god til at simulere afstrømning i et andet (opstrøms) punkt, eller at en model som er valideret mod trykniveaudata fra nogle borer ikke er nær så nøjagtig til at simulere trykniveauer i andre borer.

Tilsvarende er udtrykket en models gyldighed ofte blevet anvendt uden kvalificering med hensyn til hvor nøjagtig modellen er til givne formål. Man kan sagtens forestille sig en model, som er nøjagtig nok (og dermed gyldig) til nogle formål, men ikke til andre. Således vil der oftest stilles større nøjagtighedskrav til en model der anvendes til "hi fi" akvifersimuleringer end til modeller, der anvendes til indledende overslagsberegninger (jf. afsnit 2.4).

Det er derfor meget vigtigt at fastslå princippet om, at en model kun må karakteriseres som valideret på de områder, hvor der er gennemført valideringstest, og at der til en models gyldighed altid skal knyttes en nøjagtighed eller en usikkerhed.

Det vil eksempelvis sige, at en model, der kun er valideret mod trykniveaudata, kun har en dokumenteret gyldighed til simulering af trykniveauer, og oven i købet kun for de punkter, hvor der er valideringsdata. En sådan model er ikke dokumenteret at være gyldig også fx til simulering af minimumsvandføringer eller strømningshastigheder i grundvandet. Den kan heller ikke på forhånd antages at være lige så god til at simulere trykniveauer på andre punkter, end hvor den er valideret. Til sidstnævnte formål skal man gennemføre nogle sted-afhængige tests ligesom beskrevet i afsnit 15.4.2 ovenfor.

15.5 *ER KRAVENE TIL MODELVALIDERING UREALISTISKE I PRAKSIS?*

Ovenstående krav til modelvalidering er angivet ud fra ideelle fordringer til stringens. Hvis man sammenholdt disse fordringer med dagens praksis i grundvandsmodellering ville det formentlig være et fåtal af de eksisterende modeller, som kunne opnå betegnelsen "validerede modeller". Det betyder ikke nødvendigvis at disse modeller er værdiløse, blot at deres pålidelighed ikke er dokumenteret, og at man derfor ikke bør anvende dem som grundlag for større beslutninger. For nogle modellers vedkommende (hvor datagrundlaget eksisterer) kan der forholdsvist nemt gennemføres de relevante valideringstest.

I andre tilfælde eksisterer der ikke de relevante data til at gennemføre de ideelle valideringstest. Det kan fx dreje sig om en model, der påtænkes anvendt til at vurdere konsekvenserne af en oppumpning. Såfremt der slet ikke eksisterer data for situationer med varierende oppumpning i det pågældende opland kan det være nødvendigt at lave test på sammenlignelige nabooplunde, som beskrevet i afsnit 15.3 under differential split-sample test. Sådanne test kan aldrig blive så stærke som test på det "rigtige" opland, hvorfor konklusionerne af valideringstestene skal tages med visse forbehold. Såfremt der findes prøvepumpningsdata fra området, kan de styrke valideringsgrundlaget.

Såfremt valideringstestene viser, at modellen ikke opnår den ønskede nøjagtighed, skal man i princippet gå tilbage til Trin 2 i modelprotokollen (etablering af hydrogeologisk tolkningsmodel, jf. figur 1.2) og gentage hele processen, herunder vurdere om der er behov for yderligere feltdata. Alternativt kan man renoncere på ambitionsniveauet og eksplicit acceptere, at modellen har en mindre nøjagtighed end oprindeligt ønsket.

Selvom kravene til modelvalidering kan opfattes som ideelle og måske ikke altid nemme at opfylde i praksis, er det af afgørende betydning at være stringent og ærlig i angivelsen af en models valideringsstatus. Ved at gennemgå valideringsfasen tvinges brugeren eksplicit til at overveje og forsøge at dokumentere, hvor god en given model er, og hvor begrænsningerne i dens anvendelse findes. Stringens omkring modelvalidering er altafgørende for troværdigheden af grundvandsmodeller i praksis.

15.6 REFERENCER

- Ambroise, B., Perrin, J.L. and Reutenauer, D. (1995) Multicriterion validation of a semidistributed conceptual model of the water cycle in the Fecht Catchment (Vosges Massif, France). *Water Resources Research*, 31(6), 1467-1481.
- Anderson, M.P. and Woessner, W.W. (1992) The role of postaudit in model validation. *Advances in Water Resources*, 15, 167-173.
- Beven, K. (1996) A discussion on distributed hydrological modelling. In: *Abbott, M.B. and Refsgaard, J.C. (Eds): Distributed Hydrological Modelling*, 255-278, Kluwer Academic Publishers.
- Dee, D.P. (1995) A pragmatic approach to model validation, In: D.R. Lynch and A.M. Davies (Eds) *Quantitative skill assessment of coastal ocean models*, AGU, Washington, 1-13.
- De Marsily, G., Combes, P. and Goblet, P. (1992) Comments on 'Ground-water models cannot be validated', by L.F. Konikow and J.D. Bredehoeft, *Advances in Water Resources*, 15, 367-369.
- Flavelle, P. (1992) A quantitative measure of model validation and its potential use for regulatory purposes. *Advances in Water Resources*, 15, 5-13.
- Jayatilaka, C.J., Storm, B. and Mudgway, L.B. (1998) Simulation of water flow on irrigation bay scale with MIKE SHE. *Journal of Hydrology*, 208, 108-130.
- Konikow, L.F. (1978) Calibration of groundwater models. In: *Verification of mathematical and physical models in hydraulic engineering*, American Society of Civil Engineering, New York, 87-93.
- Konikow, L.F. and Bredehoeft, J.D. (1992) Ground-water models cannot be validated. *Advances in Water Resources*, 15, 75-83.
- Klemes, V. (1986) Operational testing of hydrological simulation models. *Hydrological Sciences Journal*, 31, 13-24.
- Larsson, A (1992) The international projects INTRACOIN, HYDROCOIN and INTRAVAL, *Advances in Water Resources*, 15, 85-87.
- Lidén, R (2000) Conceptual runoff models for material transport estimations. PdD dissertation. Report No 1028, Lund Institute of Technology, Sweden.
- Oreskes, N., Shrader-Frechette, K. and Belitz, K. (1994) Verification, validation and confirmation of numerical models in the earth sciences. *Science*, 264, 641-646.
- Refsgaard, J.C. and Knudsen, J. (1996) Operational validation and intercomparison of different types of hydrological models. *Water Resources Research*, 32(7), 2189-2202.
- Refsgaard, J.C. (1997) Parameterisation, calibration and validation of distributed hydrological models. *Journal of Hydrology*, 198, 69-97.
- Schlesinger, S., Crosbie, R.E., Gagné, R.E., Innis, G.S., Lalwani, C.S., Loch, J., Sylvester, J., Wright, R.D., Kheir, N. and Bartos, D. (1979) Terminology for model credibility. SCS Technical Committee on Model Credibility. *Simulation*, 32(3), 103-104.
- Styczen, M. (1995) Validation of pesticide leaching models. In: *Leaching models and EU registration*. Final report of the work of the Regulatory Modelling Work Group of FOCUS (Forum for the Co-ordination of pesticide fate models and their Use), DOC.4592/VI/95, European Commission.
- Tsang, C.-F., (1991) The modelling process and model validation. *Ground Water*, 29, 825-831.