Universiteit Utrecht

Sharing science,
shaping tomorrow

ORATIE

Transities in Geosimulatie

Transitions in Geosimulation

Derek Karssenberg

Universiteit Utrecht

Transities in Geosimulatie

Transitions in Geosimulation

Derek Karssenberg
Hoogleraar Computational Geography

Inaugurele rede uitgesproken bij de aanvaarding van het ambt van hoogleraar aan de faculteit Geowetenschappen van de Universiteit Utrecht, op 11 Oktober, 2023

Inaugural Lecture on the Occasion of the Acceptance of the Professorship in Computational Geography at the Faculty of Geosciences, Utrecht University, October 11, 2023

COLOFON

ISBN
978 90 6266 665 2

Uitgave
Faculty of Geosciences – Utrecht University, 2023

Foto Derek Karssenberg
Ed van Rijswijk

Foto omslag
Oppervlakkige afstroming van water, uitvoer van een hydrologisch model ontwikkeld met het LUE raamwerk (https://lue.computationalgeography.org). Ongepubliceerde data. Uitsnede van 50 km2 van een simulatie van Tirol (Oostenrijk) op een ruimtelijke resolutie van 5 m. Dieper blauwe kleuren refereren naar meer afstroming. De kaart is naar het Noorden georiënteerd; in het midden stroomt van Noord naar Zuid de Kalserbach, ten Noorden van Kals am Grossglockner. Als invoer voor het model is gebruik gemaakt van: hoogtemodel, https://www.data.gv.at/katalog/de/dataset/land-tirol_tirolgelnde; meteodata, Saha et al. (2010), doi:10.1175/2010BAMS3001.1

Surface runoff, output of a hydrological model developed with the LUE framework (https://lue.computationalgeography.org). Unpublished data. Section of 50 km^2 of a simulation of Tyrol (Austria) at a spatial resolution of 5 m. Deeper blue colours refer to more runoff. The map is oriented north; in the centre, the Kalserbach flows from north to south, north of Kals am Grossglockner. The following data were used as input to the model: elevation model, https://www.data.gv.at/katalog/de/dataset/land-tirol_tirolgelnde; meteorological data, Saha et al. (2010), doi:10.1175/2010BAMS3001.1

Grafische verzorging
C&M (10341) – Faculty of Geosciences – Utrecht University

Transities in Geosimulatie

Mijnheer de Rector Magnificus,

Geachte Hoogleraren, beste studenten, collega's, familie en vrienden,

Op welke manier probeert de wetenschap onze kennis te vergroten en indien mogelijk voorspellingen te maken van bepaalde fenomenen? Centraal staat vaak het concept van een model. Een model wordt vaak gedefinieerd als een representatie van de werkelijkheid. In de wetenschap is een belangrijk doel dat een dergelijke representatie zo veel mogelijk objectief is, dat wil zeggen vrij van subjectieve gezichtspunten. In de computationale geografie worden representaties vaak gemaakt door middel van een computer simulatie model, een nabootsing van de werkelijke geografische wereld in de computer.

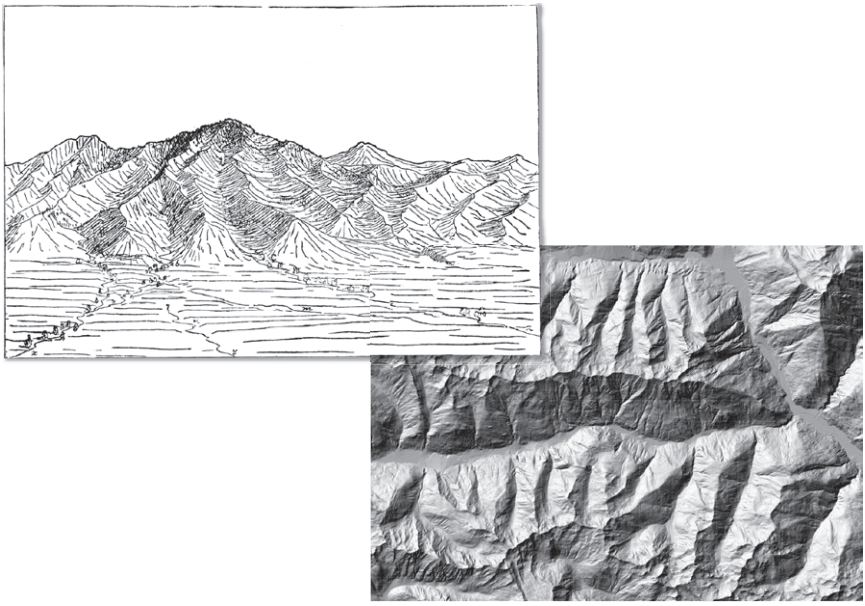
Voordat ik hier verder op in ga zou ik kort de blik willen verbreden. Laten we enkele representaties beschouwen die eerder zijn gebruikt, en soms nog steeds worden gebruikt, in de geografie of geowetenschappen. Van al deze representaties is het uitgangspunt objectiviteit. Toch is het zo dat deze representaties sterk verschillen qua uitgangspunt en methoden.

Ik begin deze korte terugblik rond 1800 en neem afbeeldingen als voorbeeld.[1] Wetenschappers maakten geïdealiseerde afbeeldingen om objecten te representeren (Figuur 1, links). Deze werden gemaakt door objecten diepgaand te analyseren om op deze wijze de universele eigenschappen te identificeren. Deze eigenschappen werden vervolgens geïdealiseerd in een visualisatie. Het interessante is dat wetenschappers nauw samenwerkten met kunstenaars; het waren vaak kunstenaars die deze tekeningen maakten, in overleg met wetenschappers. Deze benadering van natuurgetrouwe afbeeldingen kwam in de loop van de 19e eeuw onder druk te staan, met name omdat de idealisatie geen recht deed aan de natuurlijke variabiliteit.

Als antwoord hierop kwam een methode in zwang die juist niet geïdealiseerde representaties maakte van een groot aantal specifieke objecten. Dit gebeurde op een mechanische manier (Figuur 1, rechts). De idealisering van de vorige methode werd hier juist afgezworen: afbeeldingen werden gemaakt met afzwering van de wil. Dit werd enorm vereenvoudigd door de

opkomst van technieken om mechanisch afbeeldingen te maken, met name de fotografie. Deze mechanische methode voor het maken van representaties wordt nog steeds veel gebruikt. Het werkveld van de aardobservatie bijvoorbeeld maakt gebruik van beelden van de aarde, vaak op grote afstand genomen, om beschrijvingen te geven van ruimte-tijd patronen in bepaalde phenomenen.

De methode van mechanisch gemaakte representaties leverde veel informatie op maar leidde juist door het uitsluiten van de wetenschapper als de persoon die interpretatie geeft niet tot universeel begrip. Men kan natuur-



Figuur 1. Links, voorbeeld van de methode van natuurgetrouwe afbeeldingen; 'Ideal sketch to illustrate the Shifting of waterways on a slope of Planatation'. Figuur 62 in Gilbert, G. (1877), Report on the Geology of the Henry Mountains, US Geological Survey. Rechts, voorbeeld van de methode van mechanische objectiviteit; digitaal hoogtemodel van het Defereggen dal, Oostenrijk, gevisualiseerd met de 'hillshading' techniek, data van https://www.data.gv.at/katalog/de/dataset/land-tirol_tirolgelnde

lijk pogen deze beelden met de hand te classificeren, maar dit is tamelijk subjectief. Voor meer universele representaties was het nodig de structuur van de echte wereld te beschrijven, of wel de mechanismen die leiden tot bepaalde fenomenen. En dit gaat niet meer in afbeeldingen, maar met wetten, die vaak ook kunnen worden uitgedrukt in formules die de ontwikkeling van systemen kunnen beschrijven. In de geografie of geowetenschappen zijn dit vaak de wetten van Newton. Deze worden als universeel geldend gezien, waarbij aangetekend moet worden dat Newton grotendeels in het midden liet wat de achterliggende oorzaak was van zwaartekracht.[2] Naast deze fysische wetten worden er veel andere min of meer algemeen geldende wetten gebruikt als basis van formules om de echte wereld in de geowetenschappen te beschrijven.

En hiermee raken we aan het vakgebied van de computationele geografie aangezien precies deze wetten vaak (niet altijd) de basis zijn van computer simulaties.

Wat bracht de computertechnologie? Natuurlijk een rekenkracht die in de achttiende eeuw ondenkbaar was. Om een indruk te geven, de grootste supercomputer kan een triljoen – een 1 met 18 nullen – berekeningen uitvoeren per seconde.[3] Dit is gelijk aan ongeveer 1 berekening per kubieke meter water in onze oceanen. Sceptici zullen dit misschien nog tegen vinden vallen, maar het is duidelijk dat een dergelijke capaciteit enorme mogelijkheden biedt voor het maken van nieuwe representaties van het landschap: computermodellen. Er zijn veel soorten computermodellen. De focus van mijn leerstoel is op simulatiemodellen die een representatie van de wereld geven door de achterliggende mechanismen te simuleren die ten grondslag liggen aan de ontwikkeling van een geografisch systeem over de tijd.

Het basisprincipe van dit soort modellen is eenvoudig. Omdat de computer slecht kan omgaan met continue fenomenen wordt zowel de geografische ruimte als de tijd opgesplitst – gediscretiseerd – in stappen. Voor de ruimtelijke discretisatie zijn veel mogelijkheden – ik kom hier later nog op terug – maar voor de tijd worden meestal tijdstappen met vaste duur gebruikt. Een model rekent vooruit in de tijd deze tijdstappen door. Op elke stap wordt op basis van de staat van het systeem en externe invoer een berekening uitgevoerd. Deze berekening bootst de processen na die in het echt plaatsvinden in een tijdstap. De berekening resulteert in een bepaalde

nieuwe staat van het geografisch systeem, die weer invoer is voor de berekening in de volgende tijdstap. Omdat achterliggende mechanismen die het systeem sturen meestal niet veranderen, kan de berekening die elke tijdstap wordt uitgevoerd vaak hetzelfde blijven. Wat echter wel verandert is de uitvoer van die berekening: dit zorgt voor een ontwikkeling van het systeem over de tijd.

In alle sub-domeinen van de geowetenschappen worden simulatiemodellen gebruikt. Om het watersysteem te simuleren worden regen-afvoer modellen gebruikt, om te berekenen wat het landgebruik is over 10 jaar worden modellen gebruikt die de verandering in landgebruik simuleren, om blootstelling van mensen aan luchtvervuiling te bepalen worden modellen gebruikt die individuele beweging van mensen en wat ze inademen berekenen. Waarom zijn deze waardevol in de wetenschap? In de eerste plaats omdat ze bijdragen aan het testen van theorieën; ik kom hier later nog op terug. Een andere belangrijke rol van simulatiemodellen in de wetenschap is dat deze in staat zijn voorspellingen te maken van het systeem, en zo nodig voor allerlei scenario's van externe beïnvloeding. Met name deze capaciteit om te voorspellen maakt dat simulatiemodellen een enorme maatschappelijke relevantie hebben. Het weer van morgen, stikstofdepositie, zeespiegelstijging, toekomstige verspreiding van ziektes: het wordt vaak uitgerekend met simulatiemodellen. Simulatiemodellen en computationele technieken zijn dus niet meer weg te denken uit de fundamentele en toegepaste wetenschap. Sommige wetenschapsfilosofen beschouwen de computationele wetenschap zelfs als een separaat wetenschapsdomein, dat bestaat naast de geesteswetenschappen, levenswetenschappen, sociale wetenschappen, en natuurwetenschappen.[4]

Ingrediënten van modellen

Het principe van voorwaarts modelleren is universeel. Wat een specifiek model berekent echter, wordt bepaald door de gebruikte berekeningen per tijdstap, hoe deze worden uitgevoerd met een computer, en wat de invoer variabelen zijn van een model. Wat bepaalt deze vorm of structuur van een model? Een combinatie van factoren speelt een rol. Ik refereer hier naar deze factoren als *ingrediënten.*[5] De combinatie van deze ingrediënten stuurt de bouw van een model of bepaalde familie van modellen. De beschikbare

theorie over hoe het gemodelleerde systeem werkt, is het eerste ingrediënt. Het tweede ingrediënt is technologie. Dit omvat methoden die ten grondslag liggen aan de representatie van de echte wereld in de computer. Het derde ingrediënt zijn meetgegevens van het beschouwde systeem, beter bekend als 'data'. Deze kunnen op verschillende manieren worden gebruikt om het model zo goed mogelijk te laten lijken op het geografische systeem. Tot slot speelt de modelbouwer zelf en haar sociale omgeving een rol, aangezien ook psychosociale aspecten de keuzes bepalen die gemaakt worden bij het ontwikkelen van een model.

De ingrediënten zijn dus wetenschap, technologie, data, en sociologie. Mijn leerstoel beoogt de rol van deze factoren in wetenschappelijke simulatiemodellen te bestuderen en expliciet ook om het gebruik van deze ingredienten te optimaliseren. Computationele geografie op zichzelf is een wetenschap, maar het wordt uiteraard nog interessanter indien kruisbestuiving ontstaat met andere vakgebieden. Wij werken nu al vruchtbaar samen met andere wetenschappers, bijvoorbeeld binnen de hydrologie, epidemiologie, en energiewetenschappen. Deze lezing is zeker ook een uitnodiging naar collega's binnen en buiten de Universiteit Utrecht om met ons samen te werken.

Maar wat zijn de belangrijkste uitdagingen wat betreft simulatiemodellen? Ik zal u meenemen in een aantal belangrijke onderzoeksthema's gerelateerd aan de verschillende ingrediënten van modellen. Hierbij zal ik de transities beschrijven uit de titel van deze lezing.

Theorie

Laten we eerst het meest voor de hand liggende ingrediënt van modellen beschouwen: theorie. De rol van theorie kan worden begrepen als we het model beschouwen als mediator tussen theorie en data.[6] Deze mediator is nodig omdat theorie en meetgegevens (data) geen overlap vertonen. Een theorie is slechts een beschrijving, soms in formules, van algemeen geldende regels. Theorie geeft niet direct getallen die iets zeggen over wat we zouden meten in de wereld. Theorie informeert ons niet over de temperatuur van morgen. Helaas bevat meetdata ook geen theorie. Het zijn enkel getallen. Een simulatie model is dus nodig om theorie en meetdata

te verbinden. Dit kan door een model af te leiden van de theorie waardoor een model wordt verkregen dat werkt, zoals men zegt, volgens regels beschreven in de theorie. Vervolgens kan het model worden gebruikt om uitvoer te genereren, dat wil zeggen, gesimuleerde gegevens. Dit laat toe om voorspellingen te maken, bijvoorbeeld van winderosie of vegetatiegroei, gebaseerd op een wetenschappelijke theorie. Een andere manier om het model als mediator in te zetten is om het te gebruiken om theorie te testen. Door gemeten data, bijvoorbeeld diezelfde vegetatie, te vergelijken met de gemodelleerde vegetatie, kunnen we controleren of het model en daarmee de achterliggende theorie klopt.

Theorie is dus een belangrijke basis, omdat er in veel vakgebieden veel theoretische kennis is. Er zijn echter enkele beperkingen. Het omzetten van een theorie in een model is niet een puur deductieve activiteit. Het model wordt namelijk niet direct afgeleid van theorie maar er is eerder sprake van een omzetting van theorie in een model, waarbij allerlei technische hobbels moeten worden genomen, zoals het opsplitsen van de tijd in tijdstappen of het toevoegen van in werkelijkheid niet bestaande fenomenen.[7] Zelfs met deze beperkingen is het nog steeds zo dat theorie een belangrijk ingredient is van simulatie modellen. Geldt dat voor alle situaties? Helaas niet. Dit komt doordat bruikbare theorie vaak afwezig is. Voor enkele geografische systemen bestaat een goed theoretisch kader uitgedrukt in formules waarop modellen kunnen worden gebaseerd. Dit is met name zo voor systemen die beschreven worden met natuurkundige wetten, bijvoorbeeld waterstroming in het landschap. In mindere mate geldt dit voor ecologische systemen. Echter, wanneer menselijk handelen of menselijke eigenschappen worden beschouwd is direct tot modellen afleidbare theorie beperkt beschikbaar. Aangezien vrijwel alle geografische systemen een menselijke component hebben is het dus vaak slecht mogelijk een model te ontwikkelen dat geheel gebaseerd is op theorie. Een voorbeeld is het voorspellen van klimaat. Het klimaatsysteem zelf wordt hierbij gesimuleerd met van theorie afgeleide modellen. Het menselijke systeem, dat het klimaatsysteem beïnvloedt, bijvoorbeeld door de uitstoot van broeikasgassen, wordt meestal slechts gerepresenteerd door vooraf bepaalde scenario's van emissie. Hierdoor wordt de belangrijke terugkoppeling tussen klimaat en menselijk handelen beschouwd op een sterk vereenvoudigde manier. Dit voorbeeld illustreert het algemene beeld dat in veel simulatie modellen de deelsystemen bepaald door menselijk handelen slecht zijn gerepresenteerd. Dit is een belangrijke

beperking aangezien de mens zodanig veel invloed heeft op onze omgeving dat het menselijke systeem dient te worden gezien als een geïntegreerd onderdeel van het 'natuurlijke' systeem.[8]

We staan dus voor een enorme uitdaging wat betreft theorie als ingrediënt van numerieke simulatie modellen. Eenvoudig gezegd is er meer theorievorming nodig die bruikbaar is als ingrediënt van simulatiemodellen, voor een breder scala aan deelsystemen, waaronder met name ook deelsystemen waarbij menselijk handelen relevant is. Deze transitie naar betere simulatie van socio-economische deelsystemen, gekoppeld aan fysische deelsystemen, is gedeeltelijk al gaande. Recente succesvolle voorbeelden zijn de sociale hydrologie, de modellering van landgebruiksverandering, en de modellering van menselijke blootstelling aan omgevingsfactoren relevant voor gezondheid. In deze domeinen worden sociale processen die interacteren met fysisch systemen uitgedrukt in kwantitatieve formules.[9] Om dit soort modellering mogelijk te maken is nauwe samenwerking nodig tussen alpha, beta, en gamma onderzoekers. Onze Faculteit Geowetenschappen omvat al deze domeinen en is hiervoor een mooie broedplaats.

Technologie: data modellen

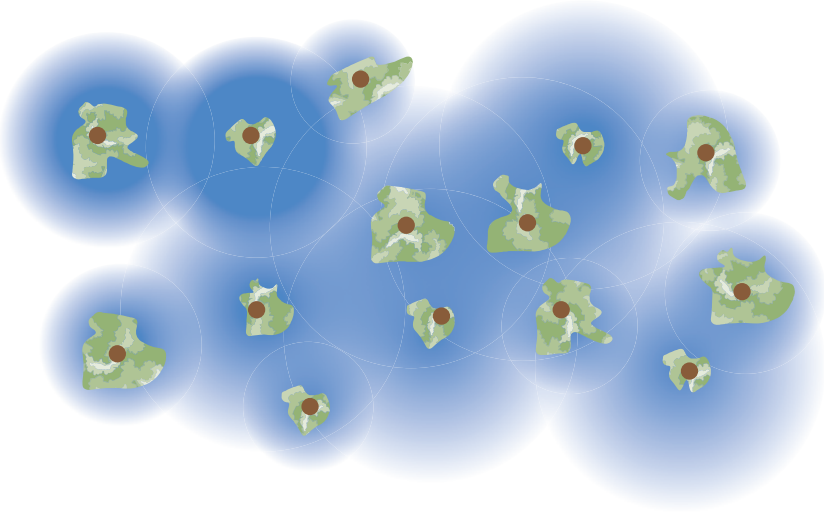
Computer technologie is een relevant ingrediënt van geosimulatie. Het bepaalt hoe conceptuele modellen van een systeem, vaak uitgedrukt in wiskundige formules en interacterende componenten, kunnen worden 'doorgerekend'. Wat is hiervoor nodig? Met name een aantal afspraken die berekeningen formaliseren zodanig dat de computer kan worden geprogrammeerd. Hiervoor bestaan modelbouw raamwerken – software specifiek bedoeld voor het maken van simulatie modellen.[10] Het te gebruiken raamwerk heeft veel gevolgen voor hoe een model wordt gebouwd, wat er mee kan worden gedaan, en wat er nodig is om het uit te breiden of te onderhouden. Raamwerken dienen een abstractieniveau te hebben dat overeenkomt met dat van de onderzoekers die modellen ontwikkelen, de modelleurs. Anders dan programmeurs die met name abstraheren in termen van IT concepten, denken modelleurs in termen van componenten of mechanismen uit het te representeren systeem. Een modelbouw raamwerk dient dus bouwstenen voor modellen aan te leveren die bepaalde mechanismen uit de echte wereld simuleren. Een bouwsteen bijvoorbeeld kan

bestaan die in één keer de hoeveelheid afstroming van water berekent over het landoppervlak. Alle technische details van een dergelijke berekening hoeft de modelbouwer niet te programmeren, want die zitten ingebakken in de bouwsteen. Voor de ontwikkeling van deze bouwstenen zijn software engineers essentieel. Zij gebruiken specialistische IT kennis om de bouwstenen zo efficiënt mogelijk te laten rekenen, bijvoorbeeld op een computer cluster. Ik kom daar zo nog op terug.

De ontwikkeling van software raamwerken voor het bouwen van simulatiemodellen is een belangrijk onderzoeksthema binnen de computationele geografie. Waar bestaan dergelijke raamwerken uit? In principe uit drie componenten. Datamodellen geven een vooraf bepaalde digitale structuur om geografische data op te slaan in de computer. Computationele functies laten toe deze data te transformeren, wat nodig is om processen door de tijd na te bootsen. Tot slot verschaft een programmeertaal toegang tot de data en de computationele functies. Deze drie componenten zijn niet op zichzelf staand. Ze moeten harmonieus samengaan. Een bepaalde representatie van gegevens in een datamodel bijvoorbeeld werkt het beste met een organisatie van computationele functies die daarop aansluit.

Ik ga hier kort in op data modellen. Het conceptualiseren van de werkelijkheid in datamodellen is zeker niet nieuw al is de huidige software context wel nieuw. In de schilderkunst en muziek bestaan ook al lang verschillende stromingen die vaak geassocieerd zijn met een bepaalde manier van representatie; iets dat wetenschappers zouden zien als datamodellen van de werkelijkheid. Onderzoek binnen onze groep op het terrein van datamodellen poogt met name twee tot nu toe verschillende representatievormen te integreren.[11] De eerste vorm is de object gebaseerde representatie, ook wel bekend als agent gebaseerde representatie. Hierbij worden natuurlijke fenomenen als objecten gerepresenteerd, elk met hun eigen eigenschappen. Deze objecten kunnen interacteren en zijn vaak mobiel. Tot nu toe worden objecten meestal gezien als punten in de geografische ruimte. Maar uiteraard kunnen ze ook een bepaald geografisch gebied omvatten. De object representatie werkt goed voor fenomenen die bestaan uit meerdere vergelijkbare onderdelen die begrensd zijn, bijvoorbeeld bomen, mensen, of instituties. De tweede vorm van representatie is een continu veld. Een continu veld heeft overal in de beschouwde geografische ruimte een waarde. Fenomenen worden weergegeven door die waarde te laten variëren in de

ruimte. Deze representatie is geschikt voor zaken die continu en niet ruimtelijk begrensd zijn, bijvoorbeeld hoogte in het landschap, temperatuur, of neerslagpatronen. Voor representatie van een geografisch systeem in de computer is het vaak wenselijk om de object gebaseerde en veld gebaseerde representatie te combineren aangezien een systeem heterogeen is. Tot voor kort echter waren veld gebaseerd en object gebaseerd modelleren gescheiden werelden, elk met specifieke datamodellen. Om deze werelden samen te brengen heeft onze groep het LUE datamodel ontwikkeld waarin de scherpe scheiding tussen objecten en velden vervalt. We doen dit door een continu veld te beschouwen als een speciaal geval van een object gebaseerde representatie: een continu veld bestaat uit één object dat het gehele studiegebied omvat. Wat we ook toevoegen is de mogelijkheid om meerdere ruimtelijke contexten of domeinen te gebruiken per object (Figuur 2). Een boom, bijvoorbeeld, is een object met eigenschappen die gerelateerd zijn aan meer-



Figuur 2. Voorbeeld van het gebruik van het LUE data model voor representatie van individuele bomen.[14] Afgebeeld zijn 14 bomen in de vorm van een kaart. Elke boom heeft drie ruimtelijke contexten (gegevenslagen) elk met een andere ruimtelijke extensie: bruin, boomstam; groen, ruimtelijke variatie in eigenschappen binnen de boomkroon; blauw, omgeving van de boom bijvoorbeeld te gebruiken om zaadverspreiding van elke boom te representeren.

dere ruimtelijke contexten. De stam kan weergegeven worden als een punt in de ruimte. Eigenschappen van de boomkroon en hoe die varieren binnen die kroon, zijn gerelateerd aan een omgrensd gebied waarbinnen de kroon bestaat. Dispersie van zaden door de kroon gebeurt over een nog weer groter gebied rondom de boom.

Een voorbeeld waar deze geïntegreerde object-veld representatie wordt gebruikt is het simuleren van het dieet van individuen in interactie met de voedselomgeving van die individuen. Personen en locaties waar eten gekocht wordt zijn hierbij de objecten, en hun omgevingen worden gesimuleerd met een veld gekoppeld aan elke object. Hiermee kan dit als een complex systeem worden gemodelleerd waarbij het dieet van personen zich aanpast aan de voedselomgeving, en vice-versa. De ruimtelijke implicatie hiervan is dat zich gebieden met gezond en minder gezond dieet ontwikkelen. Het is een abstract model, dat leidt tot beter begrip van dergelijke systemen.

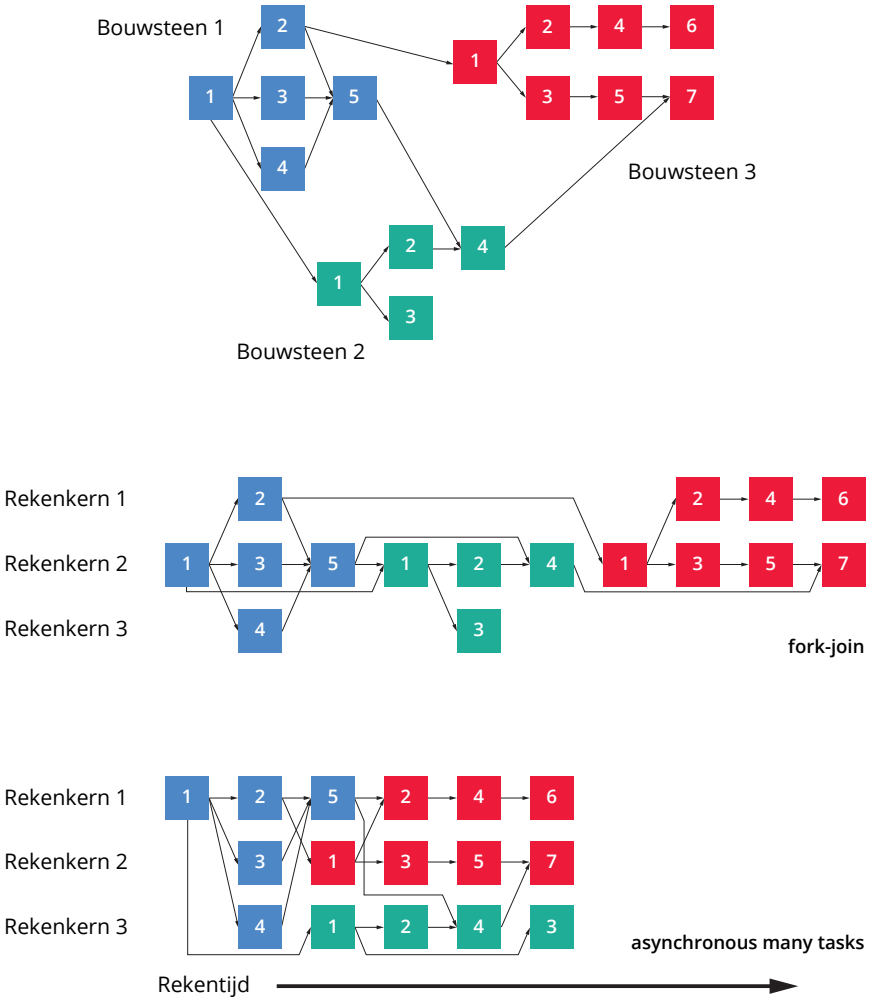
Technologie: Big Models

Er zijn dus grote uitdagingen wat betreft representatie van geografische fenomenen in de computer. Hoe zit het met de capaciteit van computers om modellen door te rekenen? Zijn ze in staat dat op voldoende snelle wijze te doen? Er is een trend in de wetenschap dat modellen steeds zwaarder worden. De term ‘Big Data’ is wellicht bekend – onvoorstelbaar grote data sets. Analoog hieraan kunnen we ook spreken van ‘Big Models’ – gigantische model berekeningen. Waardoor worden modellen zo groot? Ik noemde eerder de wens om zowel fysische als sociale fenomenen te representeren in een model. Dit soort geïntegreerd modelleren leidt tot grotere modellen, simpelweg omdat er meer modelcomponenten zijn en dus meer berekeningen. Daarnaast vereist het adresseren van belangrijke onderzoeksvragen rond klimaatverandering en gevolgen daarvan vaak modellering op mondiale schaal. De geografische omvang van modellen wordt dus groter. Tot slot heeft de verzameling van meetgegevens door middel van satellietbeelden en automatische meetapparatuur geleid tot een gigantische toename in de hoeveelheid en het detail van meetgegevens. Dergelijke gegevens kunnen worden gebruikt als invoer voor modellen. Een logisch gevolg is dat de wens bestaat om modellen ook gedetailleerder te maken.

Wat zien we echter? Wat betreft ruimtelijk en procesmatig detail blijven modellen vaak achter bij wat wenselijk en mogelijk is. Voorbeelden zijn mondiale hydrologische en landgebruiksveranderings modellen, die nog steeds rekenen op resoluties beneden die waarop processen begrepen worden en gegevens beschikbaar zijn.[12] Waardoor komt dit? Waarschijnlijk grotendeels doordat de rekencapaciteit van computers wordt onderbenut.

Er zijn enorm grote computers beschikbaar, maar het is niet zo eenvoudig die ook ten volle te benutten. Dit komt door de ingewikkelde architectuur van grote computers. De enorme verbetering van individuele rekenkernen over de laatste decennia is grotendeels tot stilstand gekomen.[13] Krachtige computers bouwt men dus tegenwoordig door enorm veel rekenkernen te combineren of meerdere computers te verbinden tot een zogenaamd computer cluster. Het aanspreken van al die rekenkernen verspreid over meerdere computers is geen sinecure. Er is speciale software technologie nodig om de berekeningen te parallelliseren over de rekenkernen of computers. Meestal wordt deze technologie pas ingezet nadat een model is ontwikkeld. Veel modellen vinden namelijk hun oorsprong enkele decennia terug, toen de mogelijkheid en de behoefte van simulaties met meer detail en over grotere gebieden nog niet bestond. Ze zijn geprogrammeerd zonder zelfs maar te denken aan de toekomstige eis van parallellisatie. Ook bij nieuwere modellen kijkt men vaak weinig vooruit wat betreft toekomstige behoeften. Naast het feit dat parallellisatie dus vaak een toevoeging achteraf is, is het al het programmeerwerk dat hiervoor wordt gedaan ook alleen maar waardevol voor dat ene enkele model. Mooi voor die ene modelbouwer, maar de wetenschappelijke gemeenschap heeft er weinig aan.

Een betere oplossing is om de capaciteit om berekeningen parallel uit te voeren in te bouwen in de eerder genoemde modelleer raamwerken. Iedereen die dat raamwerk gebruikt voor het bouwen van een model heeft dan een model dat parallel kan rekenen. De uitdaging hierbij is dat de modelbouwer een model uitdrukt door een *combinatie* van bouwstenen (meestal functies in een programma) in een modelbouw raamwerk. Elke combinatie en volgorde van bouwstenen moet het raamwerk af kunnen handelen. Om dit te begrijpen nemen we een voorbeeld van drie modelbouwstenen (Figuur 3, boven). Onder water, dat wil zeggen in het modelleerraamwerk, kan elk van die bouwstenen worden opgesplitst in kleine berekeningstaken, bijvoorbeeld per deelgebied een taak. Deze taken zijn deels afhankelijk

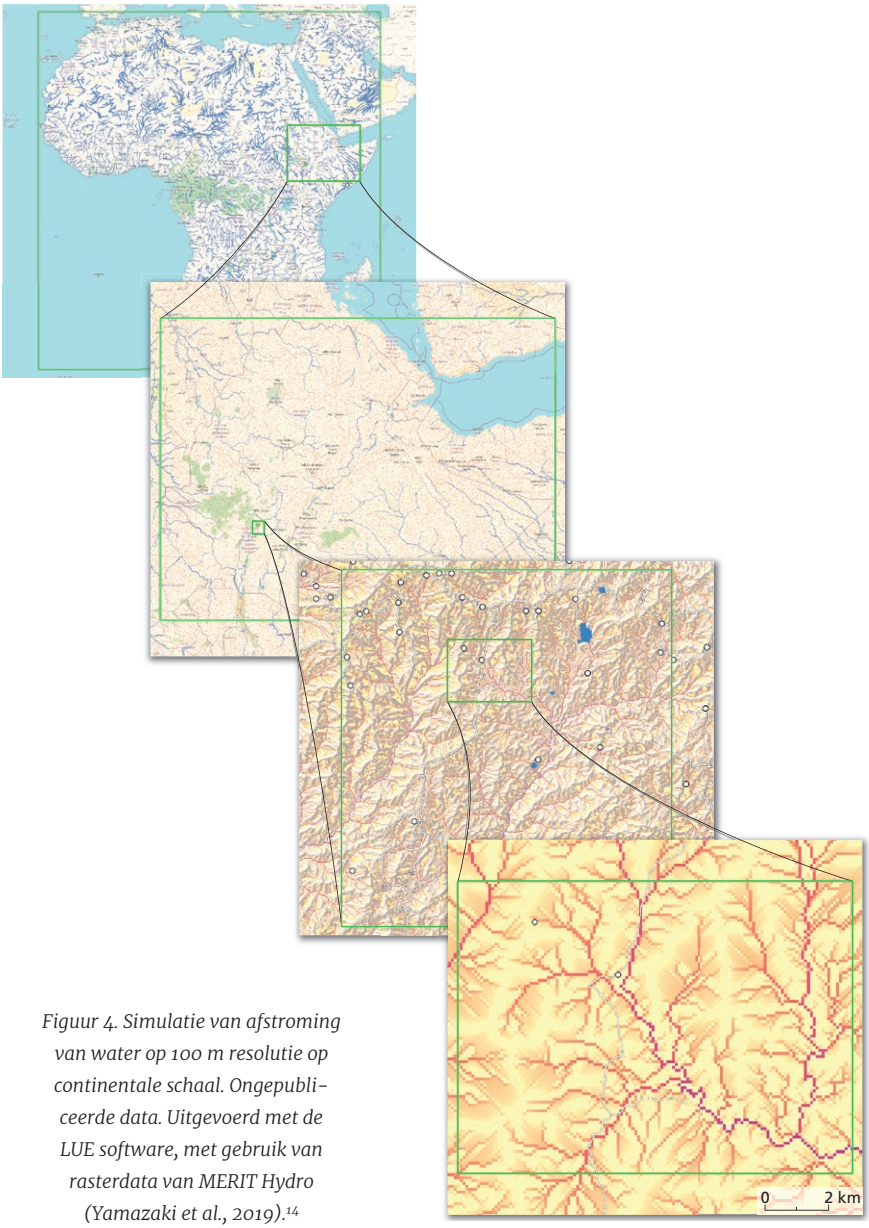


Figuur 3. Oplossingstechnieken, voorbeeld. Boven, de modelleur maakt een model door drie bouwstenen te combineren (blauw, rood, groen). Deze worden door het modelleer-raamwerk opgesplitst in taken (genummerde blokjes) die afhankelijk zijn van elkaar (pijlen). Midden, oplossingsvolgorde volgens fork-join op drie rekenkernen. Onder, idem, asynchronous many tasks (resulterend in een kortere rekentijd).

van elkaar. Als we nu de beschikking hebben over meerdere rekenkernen, in het voorbeeld drie, kunnen de berekeningen over deze kernen verspreid worden. De meest gebruikte methode zet de taken weg in de volgorde zoals de bouwstenen zijn geprogrammeerd door de modelleur: eerst de taken van de blauwe bouwsteen, dan de groene, en tot slot de rode (Figuur 3, midden). Deze methode, ook wel 'fork-join' genoemd heeft als nadeel dat niet altijd alle rekenkernen worden gebruikt doordat individuele bouwstenen niet altijd zodanig ver kunnen worden opgesplitst in taken dat alles verspreid wordt over alle rekenkernen. Vaak staan rekenkernen dus niets te doen. De methode die ons team gebruikt, ook wel asynchroon rekenen genoemd, heeft deze nadelen niet.[14] Deze methode gebruikt ook taken, maar taken worden niet in de volgorde van de bouwstenen, maar op basis van de relaties tussen de *taken* geordend en hierdoor grotendeels door elkaar uitgerekend (Figuur 3, onder). Niet synchroon dus met hoe de modelleur het heeft opgegeven. Het resultaat is een computer cluster waarbij alle rekenkernen continu aan het rekenen zijn. De rekentijd is korter. Doordat zo enorm veel taken worden aangeboden die op grotendeels willekeurige volgorde kunnen worden uitgerekend werkt het ook goed bij enorm veel rekenkernen of grote datasets.

De berekening van een standaard model omvat het doorrekenen van miljoenen bouwstenen, waarbij elke bouwsteen wordt opgesplitst in duizenden taken. Hoe is het mogelijk dat dit op de juiste manier wordt uitgevoerd, verspreid over duizenden rekenkernen die beschikbaar zijn op een computercluster? Het antwoord is het gebruik van lagen van software. De modelbouwer gebruikt een laag van software die de bouwstenen herkent. Daaronder draait een laag van software die de algoritmen kent om de bouwstenen door te rekenen, en het werk splitst in taken. Deze taken worden doorgegeven aan een meer universele bibliotheek die specifiek gemaakt is om taken uit te rekenen op meerdere rekenkernen. Hieronder draait nog allerlei software om het geheel te organiseren. Het is dus een enorm bouwwerk van componenten, waaraan honderden mensen bijdragen met behulp van een technische infrastructuur die deze samenwerking mogelijk maakt. Het zou goed zijn als geowetenschappers ook op een vergelijkbare wijze gaan samenwerken bij het ontwikkelen van modellen. Ik kom hier later nog op terug.

Concluderend is de beschikbare technologie een randvoorwaarde bij de ontwikkeling van modellen. Er is op dit terrein een enorme transitie gaande



Figuur 4. Simulatie van afstroming van water op 100 m resolutie op continentale schaal. Ongepubliceerde data. Uitgevoerd met de LUE software, met gebruik van rasterdata van MERIT Hydro (Yamazaki et al., 2019).[14]

die effect heeft op alle geo-domeinen. Het wordt mogelijk om modellen te creëren met een gedetailleerdheid die slechts enkele jaren geleden totaal ondenkbaar was. Data modellen die alle typen data kunnen representeren maken het mogelijk om heterogene systemen te modelleren. Software die clustercomputing direct beschikbaar maakt laat toe om modellen enorm groot te maken, zonder extra werk. Een voorbeeld hiervan is super hoge resolutie modelleren van hydrologie op continentale schaal waarbij resoluties van 10 tot 100 m mogelijk zijn (Figuur 4). Er liggen enorme uitdagingen om deze technologie verder te verfijnen en toe te passen in modelstudies.

Data

Verbetering van software technologie speelt dus een grote rol bij de innovatie van simulatie modellen. Hoe staat het met meetgegevens, in het kort, data? De hoeveelheid beschikbare data is de laatste jaren exponentieel gegroeid. We zouden kunnen stellen dat we zijn bedolven onder data; of misschien beter, we zijn er zelf ingesprongen. Het heeft geleid tot de aandacht voor Big Data en algoritmen, bijna dagelijks in het nieuws. In het geografische domein, worden deze data met name verzameld met behulp van automatische sensors, zoals satellietbeelden.

Wat is het gevolg van deze enorme toename in meetgegevens voor simulatie modellen? Gecombineerd met toename in rekenkracht en parallel rekenen, zoals net besproken, kan dit leiden tot een revolutie in de ontwikkeling en de rol van modellen in zowel fundamenteel als toegepast onderzoek.

Wat is er precies gaande? De hoeveelheid meetdata is zodanig groot geworden dat niet alleen wij eronder bedolven zijn, maar soms ook theoretische kennis. Het blijkt namelijk mogelijk te zijn om *alleen* gebaseerd op data voorspellende modellen te maken die het beter, en in sommige gevallen veel beter, doen dan modellen gebaseerd op theorie. Ik geef twee voorbeelden. Ten eerste de voorspelling van het weer. Recent is een puur op data gebaseerd model ontwikkeld dat het weer in bepaalde situaties beter kan voorspellen dan de meest geavanceerde theorie-gebaseerde modellen die er bestaan.[15] Het model kan dit ook nog eens doen in een fractie van de rekentijd van die theorie-gebaseerde modellen. Het tweede voorbeeld komt uit de

hydrologie. Ook hier is – al enkele jaren geleden – aangetoond dat modellen puur op data gebaseerd een betere voorspelling van rivierafvoer kunnen geven dan theorie-gebaseerde modellen.[16] Beide voorbeelden gebruiken een vorm van data gebaseerd modelleren die bekend staat als *machine learning*. Deze techniek ontwikkelt zich zeer snel.

Hoe werkt deze methode? Het principe is eenvoudig. In een eerste stap wordt een model getraind op een enorm grote data set van meetgegevens. In een tweede stap wordt dit getrainde model gebruikt voor het doen van voorspellingen. De eerste stap is met name van belang. De data set dient de sturende factoren van het systeem te omvatten, dat wil zeggen de invoer van het model, en in ieder geval ook de variabelen die voorspeld moeten worden, de uitvoer. Het trainen van het model omvat dan een berekeningsprocedure waarbij bepaalde berekeningen in het model worden geoptimaliseerd zodanig dat deze berekeningen zo goed mogelijk de uitvoer kunnen genereren, op basis van de invoer. De berekeningen zijn dus niet op processen gebaseerd. Het enige doel is om ze in staat te laten zijn de uitvoer (op basis van invoer) juist te berekenen. Hier zijn veel technieken voor. Er zijn technieken die gebaseerd zijn op een bepaald statistisch model van het systeem. Meer vrije technieken, bijvoorbeeld neurale netwerken, zijn ook gebaseerd op een bepaald model, maar laten veel meer vormen toe. Ze kunnen in principe elke relatie tussen invoer en uitvoer coderen. Het mooie van deze technieken is dat ze ook relaties in de data kunnen herkennen die misschien niet direct aan de oppervlakte liggen. Een puntenwolk van gegevens bijvoorbeeld mag misschien niet direct een relatie tonen tussen variabelen, maar een machine learning algoritme zou in staat zijn om het onderliggende mechanisme dat ten grondslag licht aan het patroon in de puntenwolk te identificeren.

Dankzij het enorme volume aan meetgegevens en de enorme capaciteit van de algoritmen zijn deze op data gebaseerde technieken potentieel enorm krachtig. Er zijn echter enkele belangrijke beperkingen. Ten eerste is deze methode alleen in staat om voorspellingen te doen voor situaties die overeenkomen met de meetgegevens. Net als een sporter zal het model alleen goed functioneren op terreinen waarop getraind is en niet daarbuiten. Een sprinter kan geen marathon lopen. Het is dus de vraag of onze data gebaseerde modellen ook nog steeds goed werken na bijvoorbeeld een grote klimaatverandering – we hebben geen gegevens uit de toekomst om het model te trainen. Ten tweede is het zeker niet altijd het geval dat er enorm

veel gegevens zijn van een bepaald gemodelleerd systeem. Van sneeuwdiktes, bijvoorbeeld, zijn slechts enkele meetstations op aarde. Erg beperkt voor een op meetdata gebaseerd model. Ten derde leveren deze modellen niet of nauwelijks begrip op van hoe een systeem werkt. Ze zijn vaak een zogenaamde 'black box'. Tot slot zijn er voor dit soort modellen nog nauwelijks procedures om onzekerheid van de uitkomsten in te schatten.

Op data gebaseerde modellen zijn dus veelbelovend maar er zijn enorme uitdagingen. Ik noem hier twee belangrijke richtingen van onderzoek. Een groot deel van de beperkingen van data gebaseerde modellen wordt veroorzaakt doordat ze in wezen 'dom' zijn. Er zit geen theorie verwerkt in het model. Een oplossing hiervoor is om zogenaamde hybride model technieken te ontwikkelen.[17] Deze technieken incorporeren theoretische kennis zonder dat de kracht en flexibiliteit van de data gebaseerde techniek verloren gaat. Een eerste benadering is om theoretische kennis direct in te bouwen in het data gebaseerde model. Een tweede manier is om het model niet alleen te trainen op meetgegevens, maar ook op gegevens die gegenereerd worden door een op theorie gebaseerd model. Het model wordt daarmee zowel gevoed door meetdata als door theoretische kennis die in het theorie-gebaseerde simulatie model zit.[18] Zelfs het gebruik van *alleen* data uit een theorie-gebaseerd model kan zeer relevant zijn omdat er surrogaat modellen mee kunnen worden gemaakt die aanzienlijk sneller rekenen dan het originele theorie gebaseerde model. Een derde methode is om theoretische kennis te extraheren uit het op meetdata gebaseerde model. Hierbij is het misschien zelfs mogelijk om dit soort modellen te gebruiken voor theorievorming.[19]

Een tweede belangrijke richting van onderzoek is het ontwikkelen van universele software en data platformen voor dit soort hybride simulaties.[20] Er bestaan al enorm veel software programma's voor theorie-gebaseerd modelleren. Evenzo bestaan er veel programma's voor puur data gebaseerd modelleren. Het is vereist om ook software te ontwikkelen voor hybride modelleren, waarbij beide paradigma worden geïntegreerd. Een ander aspect is het beter delen van data. Naast platformen voor het maken van simulaties is het essentieel dat geo-onderzoekers hun meetgegevens nog veel beter gaan delen. We moeten toe naar zogenaamde 'data spaces' waar zoveel mogelijk meetdata beschikbaar zijn, die kunnen worden gebruikt om modelbenaderingen internationaal met elkaar te vergelijken.[21]

Niettegenstaande de grote uitdagingen is het nu al zeker dat de enorme toename in beschikbare data, gecombineerd met innovatie van technieken voor het bouwen van data gebaseerde modellen, zal leiden tot een zeer grote transitie in modelbouw. De rol van data zal veel groter worden, en het is te verwachten dat automatische modelbouw met de computer een steeds grotere rol zal krijgen.

Sociologie

Het laatste en vierde aspect dat bepaalt hoe modellen vorm krijgen is de rol van de modelbouwer en haar sociale netwerk, binnen de wetenschap maar ook erbuiten. Ik beschouw eerst de rol van sociologie binnen de wetenschap. Hiervoor is het van belang eerst beter te begrijpen wat modelbouw in wezen omvat. Tijdens de bouw van een model maakt de modelleur een formele, met de computer executeerbare, beschrijving van het te modelleren landschappelijke systeem. Hierbij worden een groot aantal onderscheidingen gemaakt die vereist zijn om te komen tot de formele beschrijving.[22] Dit omvat de onderscheiding van het systeem in fenomenen, de begrenzing van het systeem en onderscheid tussen interne fenomenen en externe fenomenen die het systeem aansturen, de onderscheiding van relaties tussen fenomenen en de vorm van deze relaties, en de onderscheiding van deze fenomenen en relaties in verschillende stukken van een computer programma. Op basis waarvan dienen deze onderscheidingen gemaakt worden? Om te komen tot een kwalitatief hoogwaardig model zou het logisch zijn dat het gebeurt met name aan de hand van theorie, met als randvoorwaarden de beschikbare hardware en software, de beschikbare meetdata, en natuurlijk de onderzoeksvraag of beoogde toepassing van het model. Dit blijkt slechts ten dele het geval te zijn. Keuzes bij modelbouw blijken namelijk grotendeels te worden bepaald door het sociale netwerk waarbinnen een modelleur acteert. In de hydrologie, bijvoorbeeld, blijken keuzes grotendeels gemotiveerd door kennis en ervaring binnen het team waarin een modelleur werkt en persoonlijke ervaring en beoordeling.[23] Het blijkt dat modelleurs in de loop van de tijd sterk de neiging hebben om bepaalde benaderingen van modelbouw te internaliseren tot vaste denkpatronen, doordat ze steeds opnieuw worden geconfronteerd met diezelfde denkpatronen die bestaan in hun sociale netwerk.[24] Deze gewoontes spelen een grote rol bij modelbouw.

Is dit een probleem? Ja, dat is vrijwel zeker een probleem. Modellen spelen een belangrijke rol in het wetenschappelijk bedrijf en het is verontrustend als sociale netwerken van modelleurs en gewoontes een grote rol gaan spelen bij de ontwikkeling van modellen. Het is met name een probleem in het licht van het modelleren van extreem grote systemen. Een voorbeeld hiervan zijn geïntegreerde modellen die 'global change' beschrijven, dat wil zeggen alle interacties die een rol spelen bij mondiale verandering van klimaat, landgebruik, hydrologie, migratie, economie, inclusief beleidsvorming op dit terrein. Deze systemen omvatten zodanig veel interacties dat het enorm arbitrair wordt hoe ze worden gerepresenteerd in een model. Ze zijn dus buitengewoon vatbaar voor enigszins arbitraire keuzes gestuurd door het sociale netwerk van modelleurs.

Sociale netwerken binnen de wetenschap zijn dus deels bepalend voor de vorm die modellen aannemen. Sociologische aspecten worden nog relevanter als we de rol van modellen binnen de maatschappij beschouwen. Er zijn enorme problemen met onze natuurlijke en menselijke omgeving. Onder andere klimaatverandering, afname van biodiversiteit, schaarste van voedsel, luchtvervuiling, en migratie zijn extreem urgente problemen die onze landsgrenzen te boven gaan. Wetenschappelijke modellen worden vaak gebruikt om beleid op deze terreinen te ondersteunen. Dit heeft geleid tot een totaal nieuwe positionering van de wetenschap ten opzichte van de maatschappij. Dit wordt ook wel 'post-normal science' genoemd.[25] Post-normal science komt na 'normal science', waarbij normal science de fundamentele wetenschap en de toegepaste wetenschap omvatten. In fundamentele wetenschap is het onderzoek gescheiden van de maatschappij. In toegepaste wetenschap worden, binnen de paradigma's van de fundamentele wetenschap, bepaalde maatschappelijke vragen beantwoord. De huidige maatschappelijke problemen zijn echter zodanig groot en complex dat deze alleen kunnen worden geadresseerd in het paradigma van post-normal science. Anders dan in normal science zijn in post-normal science de vragen concreet en extreem urgent, staan waarden opeens ter discussie, en zijn er heel veel belanghebbenden. Dit soort wetenschappelijk modelleren staat niet meer naast de maatschappij maar is er onderdeel van geworden: modellen beïnvloeden het maken van beleid en de maatschappij beinvloedt de wetenschap via belanghebbenden. Modellen staan dus opeens midden in het maatschappelijk sociale systeem en de sociologie van modelbouw wordt hiermee nog relevanter.

Post-normal science is een grote transitie omdat de sociale context van wetenschappers nog belangrijker is geworden. Ik zal pogen enkele uitdagingen en oplossingsrichtingen te benoemen. Ten eerste dient de communicatie wat betreft wetenschappelijke modellen te worden herzien. Ten opzichte van de keiharde beleidsvragen die de maatschappij stelt, is wetenschappelijk modelleren, moeten we concluderen, een zachte aangelegenheid. De geografische systemen die dienen te worden gemodelleerd zijn zodanig groot, complex, en vele disciplines omvattende, dat er meerdere, niet equivalente model beschrijvingen van een systeem zijn. Het is van belang dit feit als uitgangspunt te nemen. Het is in dit licht verbazend dat juist nu de term 'digital twin' (digitale tweeling) in zwang is om modellen van grote systemen te benoemen: een model als een exacte kopie van de werkelijkheid. Aangezien er zeer vele model abstracties kunnen bestaan van dezelfde werkelijkheid is deze term verkeerd en dient niet meer gebruikt te worden.[26] Ten tweede moeten modellen transparanter worden: modelleurs dienen te beseffen dat er keuzes gemaakt worden tijdens modelbouw en deze keuzes dienen gecommuniceerd te worden.[27] Hierbij kunnen standaard modelleer procedures van nut zijn. Ten derde, om recht te doen aan de inbreng van wetenschappers uit meerdere disciplines en onderzoeksteams, belanghebbenden, en lokale actoren, dient de bouw en evaluatie van modellen meer beschouwd te worden als een groepsactiviteit.[28] Ten vierde is het vereist dat modellen echt modulair worden. Componenten kunnen hiermee eenvoudig worden uitgewisseld wat voorkomt dat onderzoeksgroepen vastgeroest komen te zitten in oude model paradigma's. Het maakt het ook eenvoudiger om modellen aan te passen aan de lokale context. In de software engineering wereld is al enorm veel kennis over het modulair en herbruikbaar maken van code, en deze kennis dient te worden ingezet in modelbouw.[29] Tevens is het van belang om geografische data internationaal te delen. Hiermee krijgen wetenschappers direct toegang tot de laatste data sets en parameter waarden die in modellen kunnen worden gebruikt. Net als modulair modelleren zal het leiden tot modellen die meer up-to-date blijven met de internationale wetenschap in plaats van gebaseerd te zijn op kennis en data in kleine onderzoeksgroepen. Tot slot is het relevant dat nieuwsgierigheid gedreven onderzoek nog steeds veel aandacht houdt; hierbij is vrijheid van denken nodig en dit gaat slecht samen met maatschappelijke relevantie. Het is bij dit onderzoek dus gewenst dat er geen enkele inbrenging is van de maatschappij in het onderwerp en de uitvoering van het onderzoek.

Conclusie

In deze lezing heb ik vier ingrediënten beschreven van geosimulatiemodellen: theorie, technologie, observationele data, en sociologie. Deze factoren bepalen de vorm van de modellen. De factoren zijn tijd en situatie afhankelijk. Dit betekent dat simulatie modellen gezien moeten worden als de best mogelijke beschrijving van een geografische systeem gegeven de beschikbare theorie, technologie, en data, en gemaakt in de context van de sociologie van een modelleur. Het is goed denkbaar dat in de toekomst andere vormen van computer modellen ontstaan. Dit is aannemelijk aangezien in het verleden en ook nu nog verschillende objectieve vormen van representatie naast elkaar bestaan in de wetenschap, zoals we aan het begin van deze lezing hebben gezien.

De ingrediënten ondergaan allen een transitie. Theorie kan steeds meer een basis vormen voor simulatiemodellen, ook voor het simuleren van menselijke deelsystemen. Technologische ontwikkeling laat het toe om extreem grote berekeningen te doen. Meetdata, gecombineerd met methoden uit de kunstmatige intelligentie, zijn zodanig rijk wat betreft detail en volume dat ze direct gebruikt kunnen worden om simulaties te bouwen. Tot slot is de sociologie van de modelleur enorm veranderd doordat modellen steeds meer worden gebruikt voor besluitvorming, waarbij meer belanghebbenden een rol kunnen gaan spelen in modelbouw. Deze transities leiden tot de mogelijkheid om zeer gedetailleerde simulaties te maken van extreem grote, geïntegreerde, systemen, waarbij de verwachting is dat modellen binnenkort grotendeels of geheel kunnen worden geïdentificeerd door de computer. De uitdagingen zijn enorm.

Naast de eerder genoemde uitdagingen bij de verschillende ingrediënten zijn er twee overkoepelende uitdagingen. Een eerste uitdaging is om het dynamische werkveld van geosimulatie te verbinden met andere onderzoeksdomeinen, zodat er kruisbestuiving plaatsvindt in wetenschappelijk onderzoek. Ten tweede is het een uitdaging om de kwaliteit van modellering te waarborgen – de nieuwe technieken vereisen nieuwe methoden om de betrouwbaarheid van voorspellingen te bepalen. Dit is relevant mede gezien het feit dat modellen een grote rol spelen in theorievorming en beleidsontwikkeling. Ik hoop deze uitdagingen aan te gaan in het kader van

mijn leerstoel in nauwe samenwerking met studenten en collega's binnen en buiten de Universiteit Utrecht.

Ik heb gezegd.

Bibliografie

Addor, N., & Melsen, L. A. (2019). Legacy, Rather Than Adequacy, Drives the Selection of Hydrological Models. *Water Resources Research*, *55*(1), 378-390. https://doi.org/10.1029/2018WR022958

Babel, L., & Vinck, D. (2022). The "sticky air method" in geodynamics. *Revue d'anthropologie Des Connaissances*, *16*(2). https://doi.org/10.4000/rac.27795

Babel, L., Vinck, D., & Karssenberg, D. (2019). Decision-making in model construction: unveiling habits. *Environmental Modelling & Software*. https://doi.org/10.1016/j.envsoft.2019.07.015

Babel, L. V, & Karssenberg, D. (2013). Hydrological models are mediating models. *Hydrological Earth System Science Discussion*, *2013*, 1053510563. https://doi.org/10.5194/hessd-10-10535-2013

Belward, A. S., & Skøien, J. O. (2015). Who launched what, when and why; trends in global land-cover observation capacity from civilian earth observation satellites. *ISPRS Journal of Photogrammetry and Remote Sensing*, *103*. https://doi.org/10.1016/j.isprsjprs.2014.03.009

Dai, Y., Shangguan, W., Wei, N., Xin, Q., Yuan, H., Zhang, S., Liu, S., Lu, X., Wang, D., & Yan, F. (2019). A review of the global soil property maps for Earth system models. *SOIL*, *5*(2). https://doi.org/10.5194/soil-5-137-2019

Daston, L., & Galison, P. (2007). *Objectivity*. Zone Books; distributed by the MIT Press.

de Bakker, M. P., de Jong, K., Schmitz, O., & Karssenberg, D. (2017). Design and demonstration of a data model to integrate agent-based and field-based modelling. *Environmental Modelling & Software*, *89*, 172-189. https://doi.org/10.1016/j.envsoft.2016.11.016

de Jong, K., & Karssenberg, D. (2019). A physical data model for spatio-temporal objects. *Environmental Modelling and Software*, *122*. https://doi.org/10.1016/j.envsoft.2019.104553

de Jong, K., Panja, D., Karssenberg, D., & van Kreveld, M. (2022). Scalability and composability of flow accumulation algorithms based on asynchronous many-tasks. *Computers & Geosciences*, *162*, 105083. https://doi.org/10.1016/j.cageo.2022.105083

de Jong, K., Panja, D., van Kreveld, M., & Karssenberg, D. (2021). An environmental modelling framework based on asynchronous many-tasks:

Scalability and usability. *Environmental Modelling & Software*, *139*, 104998. https://doi.org/10.1016/j.envsoft.2021.104998
Ducheyne, S., & van Besouw, J. (2021). Readers of the first edition of Newton's Principia on the relation between gravity, matter, and divine and natural causation: British public debates, 1687-1713. *Centaurus*, *63*(2). https://doi.org/10.1111/1600-0498.12374
Funtowicz, S. O., & Ravetz, J. R. (1993). Science for the post-normal age. *Futures*, *25*(7). https://doi.org/10.1016/0016-3287(93)90022-L
Gibbons, M., Limoges, C., & Nowotny, H. (2010). *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*. SAGE Publications Ltd.
Grayson, R. B., Moore, I. D., & McMahon, T. A. (1992). Physically based hydrologic modeling: 2. Is the concept realistic? *Water Resources Research*, *28*(10). https://doi.org/10.1029/92WR01259
Haag, D., & Kaupenjohann, M. (2001). Parameters, prediction, post-normal science and the precautionary principle – A roadmap for modelling for decision-making. *Ecological Modelling*, *144*(1). https://doi.org/10.1016/S0304-3800(01)00361-1
Horton, P., Schaefli, B., & Kauzlaric, M. (2022). Why do we have so many different hydrological models? A review based on the case of Switzerland. In *Wiley Interdisciplinary Reviews: Water* (Vol. 9, Issue 1). https://doi.org/10.1002/wat2.1574
Irrgang, C., Boers, N., Sonnewald, M., Barnes, E. A., Kadow, C., Staneva, J., & Saynisch-Wagner, J. (2021). Towards neural Earth system modelling by integrating artificial intelligence in Earth system science. In *Nature Machine Intelligence* (Vol. 3, Issue 8). https://doi.org/10.1038/s42256-021-00374-3
Karssenberg, D. (2002). The value of environmental modelling languages for building distributed hydrological models. *Hydrological Processes*, *16*(14), 2751-2766. https://doi.org/10.1002/hyp.1068
Kochiras, H. (2009). Gravity and Newton's Substance Counting Problem. *Studies in History and Philosophy of Science Part A*, *40*(3). https://doi.org/10.1016/j.shpsa.2009.07.003
Korenhof, P., Blok, V., & Kloppenburg, S. (2021). Steering Representations – Towards a Critical Understanding of Digital Twins. *Philosophy and Technology*, *34*(4). https://doi.org/10.1007/s13347-021-00484-1
Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward Improved Predictions in Ungauged

Basins: Exploiting the Power of Machine Learning. *Water Resources Research*, *55*(12). https://doi.org/10.1029/2019WR026065
Krueger, T., & Alba, R. (2022). Ontological and epistemological commitments in interdisciplinary water research: Uncertainty as an entry point for reflexion. *Frontiers in Water*, *4*. https://doi.org/10.3389/frwa.2022.1038322
Krueger, T., Page, T., Hubacek, K., Smith, L., & Hiscock, K. (2012). The role of expert opinion in environmental modelling. *Environmental Modelling and Software*, *36*. https://doi.org/10.1016/j.envsoft.2012.01.011
Latour, B. (2018). *Waar kunnen we landen? Politieke oriëntatie in het Nieuwe Klimaatregime/Down to Earth: Politics in the New Climate Regime*. Octavo Publicaties (Dutch Translation)/Polity Press (English Translation).
Lenhard, J. (2007). Computer simulation: The cooperation between experimenting and modeling. In *Philosophy of Science* (Vol. 74, Issue 2). https://doi.org/10.1086/519029
Lu, M., Schmitz, O., Vaartjes, I., & Karssenberg, D. (2019). Activity-based air pollution exposure assessment: Differences between homemakers and cycling commuters. *Health and Place*, *60*. https://doi.org/10.1016/j.healthplace.2019.102233
Magni, M., Sutanudjaja, E. H., Shen, Y., & Karssenberg, D. (2023). Global streamflow modelling using process-informed machine learning. *Journal of Hydroinformatics (Accepted for Publication)*.
Maps, F., & Record, N. R. (2020). Marine ecosystems model development should be rooted in past experiences, not anchored in old habits. *ICES Journal of Marine Science*, *77*(1). https://doi.org/10.1093/icesjms/fsz218
Masson, D., & Knutti, R. (2011). Climate model genealogy. *Geophysical Research Letters*, *38*(8). https://doi.org/10.1029/2011GL046864
Melsen, L. A. (2022). It Takes a Village to Run a Model – The Social Practices of Hydrological Modeling. *Water Resources Research*, *58*(2). https://doi.org/10.1029/2021WR030600
Montanari, A. (2015). Debates – Perspectives on socio-hydrology: Introduction. In *Water Resources Research* (Vol. 51, Issue 6). https://doi.org/10.1002/2015WR017430
Morgan, M. S., & Morrison, M. (1999). Models as mediators. In *Ideas in context*. Cambridge University Press.
Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., & Gupta, H. V. (2021). What Role Does Hydrological

Science Play in the Age of Machine Learning? In *Water Resources Research* (Vol. 57, Issue 3). https://doi.org/10.1029/2020WR028091
Oreskes, N. (2015). How earth science has become a social science. *Historical Social Research*, *40*(2). https://doi.org/10.12759/hsr.40.2015.2.246-270
Pathak, J., Subramanian, S., Harrington, P. Z., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z.-Y., Azizzadenesheli, K., Hassanzadeh, P., Kashinath, K., & Anandkumar, A. (2022). FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators. *ArXiv*, *abs/2202.11214*.
Radchuk, V., Oppel, S., Groeneveld, J., Grimm, V., & Schtickzelle, N. (2016). Simple or complex: Relative impact of data availability and model purpose on the choice of model types for population viability analyses. *Ecological Modelling*, *323*. https://doi.org/10.1016/j.ecolmodel.2015.11.022
Razavi, S., Hannah, D. M., Elshorbagy, A., Kumar, S., Marshall, L., Solomatine, D. P., Dezfuli, A., Sadegh, M., & Famiglietti, J. (2022). Coevolution of machine learning and process-based modelling to revolutionize Earth and environmental sciences: A perspective. In *Hydrological Processes* (Vol. 36, Issue 6). https://doi.org/10.1002/hyp.14596
Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195-204. https://doi.org/10.1038/s41586-019-0912-1
Rosenbloom, P. S. (2012). *On Computing: The Fourth Great Scientific Domain*. MIT Press.
Rudy, S. H., Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2021). Data-driven discovery of partial differential equations. *Science Advances*, *3*(4), e1602614. https://doi.org/10.1126/sciadv.1602614
Ruijsch, J., Verstegen, J. A., Sutanudjaja, E. H., & Karssenberg, D. (2021). Systemic change in the Rhine-Meuse basin: Quantifying and explaining parameters trends in the PCR-GLOBWB global hydrological model. *Advances in Water Resources*, *155*, 104013. https://doi.org/10.1016/j.advwatres.2021.104013
Schiavina, M., Freire, S., Carioli, A., & MacManus, K. (n.d.). *GHS-POP R2023A – GHS population grid multitemporal (1975-2030). European Commission, Joint Research Centre (JRC)*. Retrieved August 14, 2023, from http://data.europa.eu/89h/2ff68a52-5b5b-4a22-8f40-c41da8332cfe

Shen, C., Chen, X., & Laloy, E. (2021). Editorial: Broadening the Use of Machine Learning in Hydrology. In *Frontiers in Water* (Vol. 3). https://doi.org/10.3389/frwa.2021.681023

Shen, Y., Ruijsch, J., Lu, M., Sutanudjaja, E. H., & Karssenberg, D. (2022). Random forests-based error-correction of streamflow from a large-scale hydrological model: Using model state variables to estimate error terms. *Computers & Geosciences*, *159*, 105019. https://doi.org/10.1016/j.cageo.2021.105019

Sutanudjaja, E. H., van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H. C., Drost, N., van der Ent, R. J., de Graaf, I. E. M., Hoch, J. M., de Jong, K., Karssenberg, D., López López, P., Peßenteiner, S., Schmitz, O., Straatsma, M. W., Vannametee, E., Wisser, D., & Bierkens, M. F. P. (2018). PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model. *Geoscientific Model Development*, *11*(6), 2429-2453. https://doi.org/10.5194/gmd-11-2429-2018

Sutter, H. (2005). *The Free Lunch Is Over. A Fundamental Turn Toward Concurrency in Software*. http://www.gotw.ca/publications/concurrency-ddj.htm

Verstegen, J. a., Karssenberg, D., van der Hilst, F., & Faaij, A. P. C. (2014). Identifying a land use change cellular automaton by Bayesian data assimilation. *Environmental Modelling & Software*, *53*, 121-136. https://doi.org/10.1016/j.envsoft.2013.11.009

Wicke, B., van der Hilst, F., Daioglou, V., Banse, M., Beringer, T., Gerssen-Gondelach, S., Heijnen, S., Karssenberg, D., Laborde, D., Lippe, M., van Meijl, H., Nassar, A., Powell, J., Prins, A. G., Rose, S. N. K., Smeets, E. M. W., Stehfest, E., Tyner, W. E., Verstegen, J. A., … Faaij, A. P. C. (2015). Model collaboration for the improved assessment of biomass supply, demand, and impacts. *GCB Bioenergy*, *7*(3), 422-437. https://doi.org/10.1111/gcbb.12176

Winsberg, E. (2006). Models of success versus the success of models: Reliability without truth. *Synthese*, *152*(1). https://doi.org/10.1007/s11229-004-5404-6

Winsberg, E. (2010). *Science in the Age of Computer Simulation*. The University of Chicago Press.

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., & Pavelsky, T. M. (2019). MERIT Hydro: A High-Resolution Global Hydrography Map Based on Latest Topography Dataset. *Water Resources Research*, *55*(6). https://doi.org/10.1029/2019WR024873

Yokohata, T., Kinoshita, T., Sakurai, G., Pokhrel, Y., Ito, A., Okada, M., Satoh, Y., Kato, E., Nitta, T., Fujimori, S., Felfelani, F., Masaki, Y., Iizumi, T., Nishimori, M., Hanasaki, N., Takahashi, K., Yamagata, Y., & Emori, S. (2020). MIROC-INTEG-LAND version 1: A global biogeochemical land surface model with human water management, crop growth, and land-use change. *Geoscientific Model Development*, *13*(10). https://doi.org/10.5194/gmd-13-4713-2020

Noten

1 Zie Daston & Galison (2007) voor een uitvoerige beschrijving.
2 Zie Ducheyne & van Besouw (2021) en Kochiras (2009).
3 Zie https://en.wikipedia.org/wiki/Frontier_(supercomputer)
4 Rosenbloom (2012) beschrijft 'computing' als het vierde grote wetenschappelijke onderzoeksdomein.
5 Ingrediënten in de zin van vormende factoren of determinanten. Zie voor wetenschapsfilosofische behandelingen van simulatiemodellen bijvoorbeeld Winsberg (2010), Morgan en Morrison (1999), Lenhard (2007).
6 Het boek samengesteld door Morgan & Morrison (1999) geeft beschrijvingen van de mediërende rol van modellen vergelijkbaar en soms enigszins afwijkend van die hier wordt gegeven.
7 Grayson et al. (1992) legt uit waarom het modelleren op basis van fysische wetten niet altijd hoeft te leiden tot modellen die ook werkelijk het systeem beschrijven volgens dezelfde fysica. Dit wordt verder bediscussieerd in Babel & Karssenberg (2013). Haag & Kaupenjohann (2001) beschrijft het enorm grote aantal onderscheidingen (distincties) die gemaakt worden bij de bouw van een model, zelfs wanneer algemene wetten die een systeem sturen het uitgangspunt zijn. Babel en Vinck (2022) beschrijven een breed toegepaste methode om een niet bestaand fenomeen ('sticky air') toe te voegen aan geofysische simulatiemodellen om het mogelijk te maken het systeem te simuleren.
8 Oreskes (2015) bediscussieert de relatief slechte representatie van sociale processen in geosimulatie modellen. Latour (2018) beargumenteert dat menselijk handelen een vormend onderdeel is van onze omgeving en ook zo dient te worden beschouwd in wetenschap en politiek.
9 Enkele voorbeelden uit de literatuur zijn Montanari (2015) (sociale hydrologie), Wicke et al. (2015) (landgebruiksverandering), Lu et al. (2019) (menselijke blootstelling aan omgevingsfactoren).
10 Zie bijvoorbeeld Karssenberg (2002).
11 De Bakker et al. (2017) geeft de concepten van het hier beschreven data model. De Jong et al. (2019) beschrijft een software implementatie waarbij deze concepten worden gebruikt; deze software is beschikbaar op https://lue.computationalgeography.org. De door ons ontwikkelde hierbij behorende computationele functies en de programmeertaal om

modellen uit te drukken is beschikbaar als prototype software, (https://campo.computationalgeography.org).

12 Mondiale modellen die systemen aan het landoppervlak simuleren (b.v. hydrologie, landgebruik, vegetatie) hebben op het moment een ruimtelijk detail overeenkomstig met een pixelgrootte van 1 km of meer (Sutanudjaja et al., 2018; Yokohata et al., 2020). Deze maken vaak geen gebruik van invoer data die beschikbaar zijn met een ruimtelijk detail van enkele tientallen meters, bijvoorbeeld vegetatie of landgebruik (Belward & Skøien, 2015), bodem (Dai et al., 2019), bebouwing, wegen en bestraat oppervlak (https://www.openstreetmap.org), populatie dichtheid (Schiavina et al., n.d.), dit terwijl proces beschrijvingen (theorie) op deze hogere resoluties vaak wel beschikbaar zijn. Het gebruik van deze hogere resolutie data zou dus potentieel kunnen leiden tot modellen die voorspellingen geven met meer detail en zo mogelijk van hogere kwaliteit, afhankelijk van de toepassing.

13 Helaas, 'The free lunch is over'. Zie Sutter (2005).

14 Zie de Jong et al. (2021, 2022) en de software op https://lue.computationalgeography.org

15 Zie Pathak et al. (2022).

16 Zie Kratzert et al. (2019).

17 Een groot aantal reviews en editorials concluderen dat hybride simulatie de belangrijkste uitdaging is voor meetdata en AI gebaseerde simulatie binnen de geowetenschappen (Irrgang et al., 2021; Nearing et al., 2021; Razavi et al., 2022; Reichstein et al., 2019; C. Shen et al., 2021).

18 Onze groep heeft recent enkele studies gedaan waarbij theorie gebaseerde simulatie wordt geïntegreerd in data gebaseerde modellering (Magni et al., 2023; Y. Shen et al., 2022).

19 Zie bijvoorbeeld Rudy et al. (2021). Onze groep heeft enkele studies verricht waarbij uit meetdata simulatiemodellen worden geïdentificeerd die leiden tot beter begrip van mechanismen die werken in het bestudeerde systeem (Ruijsch et al., 2021; Verstegen et al., 2014).

20 De ontwikkeling van software raamwerken voor hybride simulatie is o.a. gesuggereerd door Shen et al. (2021).

21 Nearing et al. (2021) en Shen et al. (2021) benadrukken de relevantie van het delen van data sets in de hydrologie om benchmarks uit te voeren, zoals al langer gebeurt in andere AI domeinen. In een internationaal team ontwikkelt onze groep een data space rond omgevingsdata op Europees niveau (https://www.greatproject.eu).

22 Haag & Kaupenjohann (2001) geeft een uitvoerige beschrijving van de onderscheidingen (distincties) die worden gemaakt bij modelbouw en de implicaties hiervan voor modelontwikkeling.
23 Er is empirisch bewijs dat sociale processen een belangrijke sturende factor zijn bij de keuze of identificatie van model componenten of parameter waarden, o.a. in de hydrologie (Addor & Melsen, 2019; Horton et al., 2022; Melsen, 2022), geofysica (Babel & Vinck, 2022), klimaatwetenschap (Masson & Knutti, 2011), marine ecologie (Maps & Record, 2020).
24 Voor een discussie op basis van interviews afgenomen met modelleurs en sociologische theorie zie Babel et al. (2019). Winsberg (2006) beargumenteert dat modellen hun betrouwbaarheid niet alleen verkrijgen doordat ze op theorie zijn gebaseerd maar ook doordat de keuze van technieken wordt ondersteund door eerder gebruik van deze technieken.
25 Het concept post-normal science wordt geïntroduceerd in Funtowicz & Ravetz (1993). Een gerelateerde term is 'Mode 2 Science', beschreven in Gibbons et al. (2010). Krueger et al. (2012) beschrijft de rol van belanghebbenden in modelbouw in het licht van post-normal science. Hoewel Latour (2018) de term post-normal science niet gebruikt beschrijft zijn essay een vergelijkbare overgang, waarin hij stelt dat wetenschap zich dient te richten op een beschrijving van de werkelijkheid waarin de mens een geïntegreerde actor is in het aardse systeem ('les sciences de la nature-processus'); tevens beschrijft het essay dat veel belanghebbenden de ontwikkeling van wetenschappelijke kennis zullen willen beinvloeden vanwege de grote economische implicaties van wetenschappelijke kennis (bijvoorbeeld op het terrein van klimaatverandering).
26 Zie Korenhof et al. (2021) voor een kritische analyse van de term en het concept Digital Twin. In een studie op basis van een groot aantal ecologische modellen laat Radchuck et al. (2016) zien dat er veel verschillende model representaties van hetzelfde systeem mogelijk zijn, afhankelijk van de data beschikbaarheid en doelstellingen van een modelleer studie.
27 Op basis van een studie van keuze van hydrologische modellen in Zwitserland stelt Horton et al. (2022) dat de motivatie van keuzes bij modelbouw duidelijk moet worden vermeld. Vergelijkbare conclusies worden getrokken door andere studies die de sociologie van modelbouw bestuderen (Haag & Kaupenjohann, 2001; Krueger et al., 2012; Melsen,

2022). Voor de relevantie van keuzes in modelbouw zie ook Krueger & Alba (2022).

28 Participatie van belanghebbenden in modelbouw wordt bediscussieerd in Haag & Kaupenjohann (2001), Krueger et al. (2012), en Melsen (2022).

29 Zie Addor & Melsen (2019) voor de behoefte aan meer modulariteit in modelbouw.

Transitions in Geosimulation

English translation of the original Dutch text

How does science attempt to increase our knowledge and, if possible, make predictions of certain phenomena? Most often by creating a model: a representation of the real world. In science, one aims to create representations that are objective, that is, free from subjective viewpoints. In computational geography, representations are made through computer simulations, an imitation of the real geographical world in the computer.

Before discussing this in more detail, I would like to briefly broaden the scope. Let us consider some representations that have been used, and sometimes are still used, in geography or geoscience. Each of these representations use objectivity as starting point. Nevertheless, representations differ greatly in their approach.

I start this brief retrospective around 1800 and take images as an example.[30] Scientists created idealised images to represent real-world objects (Figure 1, left). These were created by analysing objects in depth to identify universal properties. These properties were then idealised into a visualisation. Interestingly, scientists worked closely with artists; it was often artists who created these drawings, in consultation with scientists. This approach of using images that were truth-to-nature came under pressure during the 19th century, particularly because the idealisation did not do justice to natural variability.

In response, a method was introduced that relied instead on non-idealised representations of a large number of specific objects. This was done in a mechanical way (Figure 1, right). The idealisation of the previous method was abandoned: images were made with renunciation of the will. This was greatly simplified by the emergence of techniques for creating images mechanically, especially photography. The mechanical method of creating representations is still widely used today; for example, in the field of Earth Observation where images are taken from Earth.

The method of mechanically created representations provided a lot of data but, precisely because of the exclusion of the scientist as the person providing interpretation, did not lead to universal understanding. One could, of course, attempt to classify these images by hand, but this is rather subjective. More universal representations were required that describe the structure of the real world, or rather the mechanisms that lead to certain phenomena. In science, this cannot be done with images, but requires laws, which can often be expressed in formulas describing temporal evolution of systems. In geography or geosciences, these are often Newton's laws. These are seen as universally valid, noting that Newton left largely open what the underlying cause of gravity was.[31] Besides these physical laws, many other more or less universally valid laws are used as the basis of formulas to describe the real world in the geosciences.

Exactly these laws and formulas are relevant to computational geography since these are often (not always as we will discuss below) the starting point of computer simulations.

What did computer technology bring? Of course, computational power that was far beyond what was possible in the 18th century. The largest supercomputer can perform a quintillion – a 1 with 18 zeros – calculations per second.[32] This is equivalent to about 1 calculation per cubic metre of water in the oceans. Sceptics may still find this disappointing, but it is clear that this capacity enables new, computerized, representations of the landscape: computer models. There are many types of computer models. The focus of my chair is on a particular branch of models: simulation models. These provide a representation of the world by mimicking the mechanisms underlying the evolution of a spatio-temporal system.

The basic principle of this kind of models is simple. As computers are poor at handling continuous phenomena, both geographical space and time are split – discretised – into steps. For spatial discretisation, there are many possibilities – I will come back to this later – but for time, time steps of fixed duration are usually used. A model computes forward in time by iterating the calculation of each time step. A calculation is performed on each step based on the state of the system and external inputs. This calculation mimics the processes that take place in the real world over a time step. The calculation results in a certain new state of the geographical

system, which in turn is input for the calculation in the next time step. Because underlying mechanisms that control the system usually do not change, the computation performed each time step can often remain the same. What does change, however, is the output of the calculation: this causes the system to evolve over time.

Simulation models are used in all sub-domains of the geosciences. To simulate the water system, rainfall-runoff models are used, to calculate what the land use will be in 10 years, models simulating land-use change are used, to determine people's exposure to air pollution, models simulating movement of individual people and their exposure to spatially varying air pollution are used. Why are these simulation models valuable in science? Firstly, because they can be used to test theory; I will come back to this later. Another important role of simulation models in science is that they are able to make predictions of the system, and if necessary for all kinds of scenarios of external influence. In particular this ability to predict makes simulation models relevant for society. Tomorrow's weather, nitrogen deposition, sea level rise, future spread of diseases: it is often calculated with simulation models. So, simulation models and computational techniques have become indispensable in basic and applied science. Some philosophers of science even consider computational science as a separate major domain of science, next to the other large domains, i.e. life sciences, social sciences, humanities, and natural sciences.[33]

Ingredients of models

The principle of simulation modelling is universal. What a specific model calculates, however, is determined by the calculations used for each time step, how they are executed with a computer, and what the input variables of a model are. What determines this structure of a model? A combination of factors is important. Here, I refer to these factors as *ingredients*.[34] A combination of these ingredients shapes the construction of a model or particular family of models. The available theory of how the modelled system works is the first ingredient. The second ingredient is technology. This includes methods to code and execute a model using a computer. The third ingredient is measurements (observations) of the system under consideration, better known as 'data'. These can be used in various ways to

make the model resemble the geographical system as closely as possible. Finally, the modeller herself and her social environment play a role, as also psychosocial aspects determine the choices made when developing a model.

So, the ingredients are theory, technology, data, and sociology. My chair aims to study the role of these factors in scientific simulation models and to optimise the use of these ingredients. Computational geography in itself is a research domain, but it obviously becomes even more interesting if cross-fertilisation with other disciplines occurs. Our team is fruitfully collaborating with other scientists, for example within hydrology, epidemiology, and energy sciences. This lecture is certainly also an invitation to colleagues inside and outside Utrecht University to cooperate with us.

But what are the main challenges as far as simulation models are concerned? Below, I will discuss key research topics related to the different ingredients of models, highlighting the transitions – rapid and ongoing innovations enabling new approaches – from the title of this lecture.

Theory

Let us first consider the most obvious ingredient of models: theory. The role of theory can be understood if we consider the model as a mediator between theory and data.[35] This mediator is necessary because theory and measurement data do not overlap. A theory is just a description, sometimes in formulas, of rules that are universally valid. Theory does not directly give numbers that say something about what we would measure in the world. Theory does not inform us about tomorrow's temperature. Unfortunately, measurement data do not contain theory either. It is just numbers. So a simulation model is needed to connect theory and measurement data. This can be done by deriving a model from theory thereby obtaining a model that follows rules described in the theory. The model then can be used to generate output, that is, simulated data. This allows us to make predictions, for example of wind erosion or vegetation growth, based on a scientific theory. Another mediating role of the model is as instrument to test theory. By comparing measured data, again for instance of vegetation,

with the modelled vegetation, we can test whether the model and thus the underlying theory used to construct the model is correct.

Theory is thus an important foundation, as there is a lot of theoretical knowledge in most research fields. However, there are some limitations. Transforming a theory into a model is not a purely deductive activity. This is because the model is not directly derived from theory but rather transformed theoretical knowledge, where the transformation requires all kinds of technical distinctions to be made, such as splitting time into time steps or adding phenomena that do not exist in the real world.[36] However, even with these limitations, theory is still an important ingredient of simulation models. Does this apply to all situations? Unfortunately this is not the case. This is because useful theory is often completely absent. For some geographical systems, an excellent theoretical framework exists expressed in formulas that can be used to deduce simulation models. This is in particular the case for systems governed by physical laws, for instance water flow in the landscape. To a lesser extent, this applies to ecological systems. Moreover, when human actions or influences are considered, theory that is transferable to models is often lacking. As most geographical systems have a human component, many models simplify or make very strong assumptions regarding the representation of certain subsystems to be modelled. An example is climate prediction. Here, the climate system itself is simulated with theory-derived models. The human system, which influences the climate system, for instance through greenhouse gas emissions, is usually only represented by predetermined scenarios of emissions. As a result, the important feedbacks between climate and human actions are represented in a simplified manner. This example illustrates the general situation that human components are often poorly represented in models. This is an important limitation since the considerable impact of humans on our environment requires consideration of the human system as an integrated part of the 'natural' system.[37]

We thus face a considerable challenge regarding theory as an ingredient of simulation models. Simply put, more theory is needed that is usable as an ingredient of simulation models, in particular for subsystems where human factors are relevant. This transition towards better theory-driven simulation of socio-economic subsystems, coupled with physical subsystems, is partly already underway. Recent successful examples

include social hydrology, the modelling of land-use change, and modelling human exposure to environmental factors relevant to health. In these fields, social processes interacting with physical systems are expressed by quantitative formulas.[38] Enabling this kind of modelling requires close collaboration between researchers from humanities, social sciences, life sciences, and natural sciences. Our Faculty of Geosciences covers these domains and can serve as incubator.

Technology: data models

Computer technology is a relevant ingredient of geosimulation. It determines how conceptual models of a system, often expressed in mathematical formulas and interacting components, can be computed. What is required for this? In particular, a set of conventions that formalise calculations enabling us to program the computer to run a model. To support the use of computers in modelling, modelling frameworks are developed which is software specifically intended for creating simulation models.[39] The framework that is used has many implications for how a model is built, how it can be applied, and what is needed to extend or maintain it. Modelling frameworks should have a level of abstraction that aligns with how modellers conceive the world. Unlike programmers who abstract mainly in terms of IT concepts, modellers think in terms of components or mechanisms from the geoscience system to be represented. A modelling framework should thus provide building blocks for models that represent certain real-world mechanisms. For example, a building block may exist that calculates the amount of runoff of water over the land surface. All the technical details of such a calculation are hidden from the modeller, as they are embedded in the building block. For the development of these building blocks then, software engineers are essential. They rely on IT expertise to make the building blocks calculate as efficiently as possible, for instance on a computer cluster. I will come back to that in a moment.

The development of software frameworks for building simulation models is an important research topic within computational geography. What are the components of these frameworks? Three components are most important. Data models provide a predetermined digital structure or framework to store geographical data in the computer. Computational functions

allow this data to be transformed, which is necessary to mimic processes over a time step in a model. Finally, a programming language provides access to the data and computational functions. These three components are not isolated but need to be integrated in a harmonious manner. For example, a particular representation of data in a data model works best with an organisation of computational functions that is aligned with the organisation of the data in the data model.

I will briefly discuss data models here. Conceptualising the real world in data models is certainly not new although the current software context is new. In painting and music, distinct styles exist that are often associated with a particular mode of representation; something that scientists would refer to as data models. In particular, research within our group in the field of data models attempts to integrate two thus far distinct forms of representation.[40] The first form is object-based representation, also known as agent-based representation. Here, natural phenomena are represented as objects, each with their own properties. These objects can interact and are often mobile. So far, objects are usually seen as points in geographical space. But of course, they can also encompass a particular geographical area. The object representation works well for phenomena consisting of multiple objects that are spatially bounded, for example trees, people, or institutions. The second form of representation is a continuous field. A continuous field has a value everywhere in the considered geographical domain. Phenomena are represented by spatial variation of this value across the spatial domain. This representation is suitable for phenomena that are spatially continuous in nature, for instance height in the landscape, temperature, or precipitation patterns. For representation of a geographical system in the computer, it is often desirable to combine the object-based and field-based representation since a system is heterogeneous; it often consists of both objects and fields. Until recently, however, field-based and object-based modelling were separate paradigms, each with specific data models. To bring these worlds together, our group has developed a data model that eliminates the sharp separation between objects and fields. We do this by considering a continuous field as a special case of an object-based representation: a continuous field consists of a single object covering the entire study area. What we also add is the possibility of using multiple spatial contexts or domains per object (Figure 2). A tree, for example, is an object with properties related to multiple spatial contexts. The trunk can be

represented as a point in space. Properties of the tree crown and how these properties vary within the crown are related to a circumscribed area within which the crown exists. Dispersal of seeds by the tree occurs over an even larger area around the tree.

An example where this integrated object-field representation is used is the simulation of the diet of individuals interacting with the food environment of those individuals. Here, individuals and food outlets are the objects, and their environments are simulated with a field linked to each object. This enables a complex system representation where the diet of individuals adapts to the food environment, and vice-versa. The spatial implication of this is that areas of healthy and less healthy diet develop. It is an abstract model leading to improved understanding of complex spatio-temporal systems.

Technology: Big Models

From the above it can be concluded that there are considerable challenges in terms of representation of geographical phenomena in computers. What about the capacity of computers to execute our models? Are they able to do so in a sufficiently fast way? There is a trend that models are getting larger and larger. The term 'Big Data' may be familiar – unimaginably large data sets. Analogously, we can speak of 'Big Models' – gigantic model computations. What is making models so large? I mentioned earlier the requirement to represent both physical and social phenomena in a model. This kind of integrated modelling leads to bigger models, simply because there are more model components and thus more calculations. In addition, addressing important research questions on climate change and its impacts often requires modelling on a global scale. Thus, the geographical scope of models increases. Finally, the collection of measurement data through satellite imagery and automated measuring devices has led to a massive increase in the amount and detail of measurement data. Such data can be used as input for models. A logical consequence is that there is a desire to make models more detailed as well.

However, what is the current status? Models often lag behind in terms of spatial and process detail with what is often desirable and possible.

Example are global hydrological and land-use change models, which still calculate at resolutions below those at which processes are understood and data are available.[41] Why is this? Probably largely because the computational capacity of computers is underutilised.

Very large computers are available, but it is not so straightforward to make full use of them. This is because of the complicated architecture of large computers. The speed-up of individual computing cores over the last few decades has largely come to a halt.[42] So nowadays, powerful computers are built by combining a large number of computing cores or by connecting several computers into a so-called computer cluster. Addressing all those compute cores distributed across multiple computers, however, is challenging. Special software technology is needed to parallelise the computations across the compute cores or computers. Usually, this technology is deployed only after a model has been developed. Indeed, many models trace their origins back several decades, when the possibility and need for simulations with more detail and over larger areas did not exist. They were programmed without considering the future requirement of parallelisation. The same holds often for more recent models. Parallelisation is mostly added afterwards the original coding phase, in an ad-hoc manner. This leads to an improved model, but it does not contribute to development of other models as the parallelisation is model specific.

A better solution is to build the ability to perform calculations in parallel into the modelling frameworks mentioned earlier, that is, under the hood. Anyone who uses the framework to build a model then has a model that will execute calculations in parallel. The challenge in designing such a framework is that it needs to be able to parallelize any combination and sequence of model building blocks (usually functions) used in a model. To understand this, let's take an example of three model building blocks (Figure 3, top). Inside the modelling framework, each of these building blocks can be split into small computational tasks, for instance one task per subarea. These tasks will be interdependent. If we now have access to several computational cores, three in the example, the computations can be spread across these cores. The most commonly used method executes the tasks in the same order as how they were coded by the modeller: first the tasks of the blue building block, then the green, and finally the red (Figure 3, centre). This method, also known as fork-join, has the disadvantage that

often a number of computational cores remain unused. This is because the granularity of the tasks is often too coarse to enable distribution of tasks across all cores. An alternative method introduced into geosimulation by our team, also called asynchronous many tasks, does not have these drawbacks.[43] This method also uses tasks, but tasks of different building blocks are ordered based on the relationships between the tasks, which may be different from the ordering of the building blocks specified by the modeller (Figure 3, bottom). The result is a computer cluster where all computational cores are continuously busy doing calculations. The computation time is shorter. Because a large number of tasks is created and can be computed in a largely random order it also works well with extremely large numbers of computational cores or large data sets.

Executing a standard model involves computations of millions of building blocks (if not more), where each building block is broken down into thousands of tasks. How can a single software tool manage such a task? The answer is in the use of layers of software. The model builder uses a layer of software that recognises the building blocks. Below this layer runs software that knows the algorithms to compute building blocks, and how to split the work into tasks. These tasks are passed to a more universal layer developed specifically for computation of tasks on multiple compute cores. The solution depends on a large set of interacting components, with hundreds of people contributing to develop and maintain these components. This level of cooperation between software engineers is largely unknown of in the scientific world of model builders, and it would be beneficial if a similar community-based approach would be used by model builders just like software engineers do. I will come back to this below.

In conclusion, technology enables and constrains model development. A large transition is taking place in this field, affecting all geo-domains. Models can be constructed now with a level of detail that was totally unthinkable of only a few years ago. Data models that can represent all types of data enable modelling of heterogeneous systems. Software for cluster computing allows models to be made massive in size, without extra work. An example of this is high-resolution modelling of hydrology at the continental scale where resolutions of 10 to 100 m are possible (Figure 4). Considerable challenges remain to make this technology more sophisticated and to test it in modelling studies.

Data

Software technology and its innovation over time has thus played a major role in shaping simulation models. What about measurement data, in short, data? The amount of available data, mostly measurement data, has grown exponentially in recent years. It has led to the focus on Big Data and Artificial Intelligence, almost daily in the news. In the geoscience domain, this data is mainly collected using automatic sensors, for instance those used for collecting remote sensing imagery.

What is the consequence of the considerable increase in measurement data for numerical simulation models? Combined with a jump in computing power and parallel computing, as just discussed, this will lead to a revolution in the development and role of models in both basic and applied research. What is going on? The amount of measurement data has become so large that it is often more informative than theory as source for configuring our models. Indeed, it turns out that predictive models based *only* on data do better, and in some cases much better, than models based on theory, even for very complex, spatio-temporal systems. I give two examples. First, weather prediction. Recently, a purely measurement data-based model has been developed that can predict weather in certain situations better than the most sophisticated theory-based models out there.[44] The model can also do this in a fraction of the computing time used by theory-based models of the same system. The second example comes from hydrology where it has been shown – several years ago already – that purely data-based models can give better prediction of river discharge than theory-based models.[45] Both examples use a form of data-based modelling known as *machine learning*. This technique is developing very rapidly.

What is the approach of data-based modelling? The principle is simple. In the first step, a model is trained on a large data set of measurement data. In the second step, this trained model is used to make predictions. The first step is particularly important. The data set needs to include the driving factors of the system, that is, the input to the model, and at least also the variables to be predicted, the output. Training the model then involves a computational procedure in which certain calculations in the model are optimised to an extent that these calculations can reproduce the output, based on the input. Thus, the calculations are not theory-based, at

least not theory from geosciences. The only goal is to make them capable of calculating the output (based on inputs) correctly. There are many techniques to do so. One is relying on particular statistical models of the system. More flexible techniques, for instance neural networks, are also based on a particular model, but allow many more forms. They can basically encode any relationship between input and output. The beauty of these techniques is that they can also recognise relationships in the data that may not be apparent. For example, a scatterplot of data may – based on visual interpretation – not show any relationship between variables, but a machine learning algorithm could identify the mechanism that produced the pattern in the data.

Owing to the large volume of available measurement data and this enormous capability of algorithms, these data-based techniques are potentially extremely powerful. However, there are important limitations of the approach. One is that data-based approaches can only predict in contexts that match the measured data used for training a model. Just like an athlete, the model will only work well in settings that have been trained on. A sprinter will not perform well on a marathon. Thus, it is questionable whether certain data-based models will still work well after, say, a major climate change – we don't have data from the future that can be used for training the models. Second, it is certainly not always the case that there is sufficient data for training. For snow thickness, for example, there are only a few monitoring stations on Earth that measure the equivalent amount of water stored in a snowpack. This would be limited information for training a data-based model. Third, data-based models provide little understanding of how a system works. They are often so-called 'black box' models. Finally, for these types of models, there are still hardly any procedures to estimate uncertainty of the outcomes.

Data-based models are thus promising but there are important challenges. There are two main directions of research. Most limitations of data-based models are due to their lack of intelligence; that is, there is no theory embedded in the model. One solution is to develop so-called hybrid modelling techniques.[46] These techniques incorporate theoretical knowledge without losing the power and flexibility of the data-based approach. A first method is to incorporate theoretical knowledge directly into the data-based model. A second approach is to train the model not only

on measured data, but also on data generated by a theory-based simulation model, which implies the model is trained both on measurement data and theoretical knowledge contained in the theory-based simulation model.[47] Even relying *solely* on data from a theory-based model in the training phase can be interesting as it results in surrogate models that may compute faster than the original theory-based model. A third method is to extract theoretical knowledge from the data-based model. Here, it may even be possible to use this type of model for theory building, i.e. scientific discovery.[48]

A second important direction of research is to develop universal software and data platforms for hybrid simulation.[49] A number of modelling frameworks already exist for theory-based modelling. Similarly, many tools exist for purely data-based modelling. It is required to also develop tools for hybrid modelling, integrating both paradigms. Besides platforms for building simulations, it is essential that geoscientists start sharing their measurement data in effective and efficient manners. We need to move towards 'data spaces' providing data that can be used to compare modelling approaches.[50]

Notwithstanding the challenges, it is certain that the increase in available data, combined with innovation in techniques for building data-based models, will lead to a transition in modelling towards more data-based approaches. It can be expected that automated computer-based modelling will become important, if not a standard approach.

Sociology

The final and fourth aspect that shapes models is the modeler herself and her social network within as well as outside the scientific community. I first discuss the role of sociology within the scientific community. For this, it is important to better understand what model building comprises. When constructing a model, the modeller creates a formal, computer-executable description of the geographical system. This involves a large number of distinctions required to arrive at the formal description.[51] These include the distinction of the system into phenomena, the distinction between phenomena internal to the system and those driving the system

from outside, the distinction of relationships between phenomena and the form of these relationships, and the distinction of these phenomena and relationships into different modules of a computer program. How should these distinctions be made? Of course based on theory, enabled or constrained by the available hardware and software, the available measurement data, and the research question or intended application of the model. This turns out to be only partially the case: choices in model building are largely determined by the social network within which a modeller acts. In hydrology, for example, choices appear to be largely motivated by knowledge and experience within the team in which a modeller works and personal experience and judgement.[52] Over time, modellers have a strong tendency to internalise certain modelling approaches into particular patterns of thought, as they are repeatedly confronted with the same patterns existing in their social network.[53] These habits play a major role in modelling.

Is this a problem? Yes, it is almost certainly a problem. Models play an important role in science, and it is worrying if social networks of modellers and habits play a major role in model identification. It is in particular a problem when modelling extremely large systems. An example is integrated models that describe global change, i.e. all interactions involved in global change in climate, land use, hydrology, migration, economics, including policy making. The number of interactions in such a system is large and many configurations of models are possible to represent these systems. The building of these models is thus very susceptible to possibly arbitrary choices driven by the social network of modellers.

Social networks within science thus partly determine the form models take. Sociological aspects become even more relevant when we consider the role of models within society. As we all know we have caused and are threatened by several environmental issues with implications at local and global scale. Climate change, biodiversity decline, food scarcity, air pollution, and migration, among others, are extremely pressing problems that transcend our national borders. Scientific models are often used to support policy-making in these areas. This has led to a totally new positioning of science relative to society, referred to as 'post-normal science'.[54] Post-normal science emerged after 'normal science', with normal science encompassing basic science and applied science. In basic science, research is separated

from society. In applied science, still within the paradigms of basic science, certain societal questions are answered. However, today's environmental problems are so large and complex that they can only be addressed in the paradigm of post-normal science. Unlike in normal science, in post-normal science the questions are concrete and extremely urgent, values are open for discussion, and there are many stakeholders. This kind of scientific modelling has become part of society: models influence policy-making, and society influences modelling via stakeholders. So, certain models are suddenly essential to society, and the sociology of modelling thus becomes even more relevant as it involves social networks outside the scientific community.

Post-normal science is a major transition because the social context of scientists has become more relevant. I will attempt to identify some challenges and directions for solutions. First, communication regarding scientific models needs to be revised. Relative to the rock-hard policy questions posed by society, scientific modelling is, we must conclude, a soft business. The size, complexity, and heterogeneity of geographical systems considered in post-normal science is so enormous that multiple, non-equivalent, model descriptions of a system exist. This should be one of the starting points when communicating model results. In this light, it is surprising that rather recently the term digital twin has become in vogue to refer to models of large systems: a model as an exact copy of reality. Since there can be a large number of model abstractions of the same reality, this term is not suitable and should be abandoned.[55] Second, models need to become more transparent: modellers themselves need to realise that choices were made during modelling and these need to be communicated.[56] Third, to do justice to the input of scientists from multiple disciplines, stakeholders, and local actors, model building needs to be regarded as a group activity.[57] Fourth, it requires that models become truly modular such that components can be interchanged preventing research groups from getting stuck in old model paradigms. It also enables to adapt models to local contexts. There is already considerable knowledge in the software engineering domain about making code modular and reusable, and this knowledge can often be transferred to modelling.[58] It is also important to share geographical data. This will give scientists instant access to the latest data sets and parameter values that can be used in models. Like modular modelling, it will lead to models that stay more up-to-date with

international science instead of being based on knowledge and data in individual research groups. Finally, it is relevant that curiosity-driven research remains to exist; it flourishes best when executed with limited influence from society and therefore needs to stay as a component of our research outside post-normal science.

Conclusion

In this lecture, I described four ingredients of geosimulation models: theory, technology, data, and sociology. These factors determine the shape of models used today in the geosciences. The factors are time and situation dependent. This implies models need to be considered as the best possible description of a geographical system, given the available theory, technology, and data, and constructed in the context of a modeller's social network. It is quite conceivable that in the future other forms of computer models will emerge. This is plausible as multiple objective forms of representation have coexisted in science in the past and also today, as we saw at the beginning of this lecture.

The ingredients are all undergoing a transition. Theory can increasingly form a basis for simulation models, including the simulation of human subsystems. Technological development enables extremely large computations. Observational data, combined with methods from artificial intelligence, are so rich in terms of detail and volume that they can be used directly to build simulations. The sociology of the modeller has changed enormously as models are increasingly used for decision-making, with more stakeholders playing a role in modelling. These transitions lead to the ability to build highly detailed simulations of extremely large, integrated, systems, where it can be expected that model building can soon be partly automated. The challenges are enormous. Besides the previously mentioned challenges related to the different ingredients of modelling, there are two overarching challenges. One is to connect the rapidly evolving field of computational geography with other research domains such that these benefit from the innovations. The second challenge is to ensure the quality of modelling. New approaches to modelling require new methods to determine the reliability of predictions. This is relevant also because models play a major role in theory building and policy making. I hope to

address these challenges as part of my chair in close collaboration with students and colleagues inside and outside Utrecht University.

Figure captions

Figure 1 (page 4). Left, example of the method of truth-to-nature using idealised representations; 'Ideal sketch to illustrate the Shifting of waterways on a slope of Planatation'. Figure 62 in Gilbert, G. (1877), Report on the Geology of the Henry Mountains, US Geological Survey. Right, example of mechanical objectivity; digital elevation model of the Defereggen valley, Austria, visualised using the 'hillshading' technique, data from https://www.data.gv.at/katalog/de/dataset/land-tirol_tirolgelnde

Figure 2 (page 11). Example of the application of the LUE data model for representation of individual trees.[14] Shown are 14 trees as a map. Each tree has three spatial contexts (data layers) each with a different spatial extension: black, tree trunk; green, spatial variation in properties within the tree crown; blue, environment of the tree, for instance to be used to represent seed dispersal from each tree.

Figure 3 (page 14). Solution techniques, example. Top, the modeller combines three building blocks (blue, red, green) for building a model. These are split into interdependent (arrows) tasks (numbered) by the modelling framework. Centre, execution following the fork-join approach running on three computational cores. Bottom, idem, asynchronous many tasks approach (resulting in shorter computation time).

Figure 4 (page 16). Simulation of water runoff at 100 m resolution on continental scale. Unpublished data. Model created and executed with the LUE software, using rasterdata from MERIT Hydro (Yamazaki et al. 2019).[14]

Notes

30 Refer to Daston & Galison (2007) for a detailed description.
31 Ducheyne & van Besouw (2021) and Kochiras (2009).
32 https://en.wikipedia.org/wiki/Frontier_(supercomputer)
33 Rosenbloom (2012) describes 'computing' as the fourth major domain of scientific research.
34 Ingredients in the sense of formative factors or determinants. For philosophical accounts of simulation models, refer to, for example, Winsberg (2010), Morgan and Morrison (1999), Lenhard (2007)
35 The book compiled by Morgan & Morrison (1999) gives descriptions of the mediating role of models similar and sometimes slightly different from those given here.
36 Grayson et al. (1992) explains why modelling based on physical laws not necessarily leads to models that actually describe the system according to the same physics. This is also discussed in Babel & Karssenberg (2013). Haag & Kaupenjohann (2001) describe the large number of distinctions made when constructing a model, even when general laws governing a system are the starting point. Babel and Vinck (2022) describe a widely used method of adding a non-existing phenomenon ('sticky air') to a certain group of geophysical simulation models to enable simulation of the system.
37 Oreskes (2015) discusses the relatively poor representation of social processes in geosimulation models. Latour (2018) argues that human actions are a formative part of our environment and should also be considered as such in science and politics.
38 Some examples from the literature include Montanari (2015) (social hydrology), Wicke et al. (2015) (land use change), Lu et al. (2019) (human exposure to environmental factors).
39 See, for example, Karssenberg (2002).
40 De Bakker et al. (2017) provides the concepts of the data model described here. De Jong et al. (2019) describes a software implementation using these concepts; this software is available at https://lue.computationalgeography.org. The associated computational functions and programming language we developed to express models is available as prototype software at https://campo.computationalgeography.org.

41 Global models simulating land surface systems (e.g. hydrology, land use, vegetation) currently have a spatial detail corresponding to a pixel size of 1 km or larger (Sutanudjaja et al., 2018; Yokohata et al., 2020). These often do not use input data available at a spatial detail down to tens of metres, e.g. vegetation or land use (Belward & Skøien, 2015), soil (Dai et al., 2019), built-up areas, roads and paved surfaces (https://www.openstreetmap.org), population density (Schiavina et al., n.d.), while process descriptions at these higher resolutions are often available. The use of this higher resolution data could potentially lead to models that provide predictions with more detail and possibly of higher quality, depending on the application.

42 Unfortunately, 'The free lunch is over'. See Sutter (2005).

43 Refer to de Jong et al. (2021, 2022) and https://lue.computationalgeography.org

44 Refer to Pathak et al. (2022).

45 Refer to Kratzert et al. (2019).

46 A large number of reviews and editorials describe hybrid simulation as the main challenge for data and AI-based simulation within the geosciences (Irrgang et al, 2021; Nearing et al, 2021; Razavi et al, 2022; Reichstein et al, 2019; C. Shen et al, 2021).

47 Our group has recently conducted studies integrating theory-based simulation and data-based modelling (Magni et al., 2023; Y. Shen et al., 2022).

48 See, for example, Rudy et al. (2021). Our group has conducted studies identifying simulation models from measured data that lead to better understanding of mechanisms operating in the studied system (Ruijsch et al., 2021; Verstegen et al., 2014).

49 The development of software frameworks for hybrid simulation has been suggested, among others, by Shen et al. (2021).

50 Nearing et al. (2021) and Shen et al. (2021) highlight the relevance of sharing data sets in hydrology to perform benchmarks, as has long been done in other AI domains. Our group is participating in a team developing a data space for environmental data at the European level (https://www.greatproject.eu).

51 Haag & Kaupenjohann (2001) provides a detailed description of the distinctions made in modelling and the implications for model development.

52 There is empirical evidence that social processes are an important factor in the choice or identification of model components or parameter values, in the domain of hydrology (Addor & Melsen, 2019; Horton et al., 2022; Melsen, 2022), geophysics (Babel & Vinck, 2022), climate science (Masson & Knutti, 2011), marine ecology (Maps & Record, 2020).

53 For a discussion based on interviews conducted with modellers and sociological theory refer to Babel et al. (2019). Winsberg (2006) argued that models obtain their credibility not only because they are theory-based but also because the choice of techniques is supported by previous use of these techniques.

54 The concept of post-normal science is introduced in Funtowicz & Ravetz (1993). A related term is 'Mode 2 Science', described in Gibbons et al. (2010). Krueger et al. (2012) describes the role of stakeholders in modelling in light of post-normal science. Although Latour (2018) does not use the term post-normal science, his essay describes a similar transition, arguing that science should evolve towards a description of reality in which humans are an integrated actor in the earth system ('les sciences de la nature-processus'); also, the essay states that many stakeholders aim to influence the development of scientific knowledge because of the large economic implications of scientific knowledge (e.g. in the field of climate change).

55 Refer to Korenhof et al. (2021) for a critical analysis of the term and concept of Digital Twin. In a study based on a large number of ecological models, Radchuk et al. (2016) show that many different model representations of the same system are possible, depending on the data availability and objectives of a modelling exercise.

56 Based on a case study on hydrological models in Switzerland, Horton et al. states (2022) that the motivation of choices in modelling should be clearly communicated. Similar conclusions are drawn by other studies on the sociology of modelling (Haag & Kaupenjohann, 2001; Krueger et al., 2012; Melsen, 2022). For the relevance of choices in modelling see also Krueger & Alba (2022).

57 Stakeholder participation in modelling is discussed in Haag & Kaupenjohann (2001), Krueger et al. (2012), and Melsen (2022).

58 See Addor & Melsen (2019) on the need for modularity in model building.

Derek Karssenberg (1968) is hoogleraar ‘Computational Geography’ aan de Universiteit Utrecht, Faculteit Geowetenschappen, Vakgroep Fysische Geografie. Hij studeerde Fysische Geografie en promoveerde in 2002 aan de Universiteit Utrecht. Zijn onderzoek richt zich op het ontwikkelen van methoden en software voor het modelleren van geografische systemen en het gebruik van deze methoden binnen een breed spectrum van onderzoeksterreinen, waaronder hydrologie, ecologie, en epidemiologie. Zijn onderzoeksgroep bestaat uit wetenschappelijke programmeurs en data scientists, waarbij samen wordt gewerkt met experts in verschillende domeinen van de geografie en gerelateerde onderzoeksdisciplines, waaronder ook sociale wetenschappen. Hij heeft, samen met zijn onderzoeksgroep, de ontwikkeling geïnitieerd van meerdere open source software systemen voor geosimulatie, die de basis vormen van een aantal grote modellen ontwikkeld aan universiteiten en onderzoeksinstituten. Hij was hoofredacteur van het tijdschrift Computers & Geosciences en is redacteur van het tijdschrift Environmental Modelling & Software. Karssenberg draagt bij aan de innovatie van onderwijs, met name op het terrein van veldwerk, eLearning, en afstandsonderwijs. Daarnaast was hij één van de initiators van de Applied Data Science MSc opleiding aan de Universiteit Utrecht.