



Planbureau voor de Leefomgeving

PBL-Notitie

De ecologische kwaliteit van Nederlands oppervlaktewater: een vergelijking van drie modelleringstechnieken

H. Visser
contact: hans.visser@pbl.nl

Datum: 31-1-2013
Publicatienummer: 500208002

Woord vooraf

Hierbij treft u aan de uitgewerkte resultaten van de 29 EKR-datasets, zoals aangeleverd door Royal Haskoning (Niels Evers) en gemodelleerd door Hans Visser (regressiebomen, PBL), Ton Schomaker (neurale netwerken, Royal Haskoning) en Arie de Niet (neurale netwerken of PUNNs, Witteveen+Bos). Implementatie in de KRW-verkenner wordt uitgevoerd door Erwin Meijers (Deltares).

Niet over alle punten in deze notitie bestaat overeenstemming. Waar dat het geval is, heb ik de verschillende meningen weergegeven.

Inhoud

1	Aanleiding...	3
2	Voorspel-performance drie modellen...	4
3	Gebruik van expert judgment bij modellering...	9
4	Staan modellen een biologische interpretatie toe?...	10
5	Zijn alle wateren/EKRs even goed te voorspellen?...	13
6	Rol van ingeschatte EKR-deelmaatlat-waarden...	14
7	Samenvatting en implementatie in de KRW-verkenner...	15
	Appendix A Voorspellingen voor de validatieperiode in detail...	18

1 Aanleiding

In 2009 heeft Royal Haskoning een rapport opgesteld over het verbeteren van datasets en de afleiding van ecologische rekenregels voor de KRW-verkenner. Voorspelling-modellen werden afgeleid op basis van neurale netwerken (Royal Haskoning) en regressiebomen (PBL). Deze modellen werden getest op 29 EKR-deelmaatlatten, verspreid over 8 watertypen. Zie rapport nr. 9T6271, 9 februari 2009. Over de benadering van regressiebomen heeft het PBL een rapport gepubliceerd in 2008 (“Stuurfactoren voor de ecologische kwaliteit van regionaal oppervlaktewater”).

In 2012 heeft een update plaatsgevonden van dezelfde 29 EKR-deelmaatlatten. Datasets zijn uitgebreid en opgeschoond door Royal Haskoning. Elk van de 29 datasets is nu bij benadering 200 wateren groot. Stuurvariabelen bij elk van de 29 maatlatten zijn identiek gebleven aan die in 2009.

In deze notitie worden drie methoden vergeleken die voorspellingen kunnen genereren voor elk van deze 29 EKR-deelmaatlatten: regressiebomen (PBL), neurale netwerken (Royal Haskoning) en neurale netwerken of PUNNs (Witteveen+Bos). PUNN staat voor Product Unit Neural Network.

Deze drie methoden worden geëvalueerd op drie aspecten: voorspel-performance (hoofdstuk 2), het wel of niet toelaten van biologisch onrealistische relaties (hoofdstuk 3), en biologisch interpreteerbaarheid van de modellen (hoofdstuk 4). In hoofdstuk 5 wordt bekeken of er ook watertypen/EKR's zijn die door geen van de modellen bevredigend voorspeld kunnen worden. Hoofdstuk 6 laat zien in hoeverre de resultaten beïnvloed worden door inschattingen van EKR-waarden. Tenslotte worden in hoofdstuk 7 de resultaten samengevat en wordt bekeken wat de software-consequenties zijn voor de KRW-verkenner.

2 Voorspel-performance drie modellen

2.1 Vergelijkbaarheid van voorspellingen in de validatie-periode

Bij aanvang van het project is afgesproken dat Royal Haskoning bij de levering van de 29 EKR-bestanden ook een opdeling zou maken in de verhouding 80% (de kalibratie-set) en 20% (de validatie-set). Alle modellen worden dan geschat op de kalibratieset, zonder verder gebruik te maken van de informatie in de validatie-set. Een vergelijking van methodes is dan eerlijk te maken door de voorspel-performance te evalueren voor elk van de 29 datasets, gerekend over alle validatie-sets. Voor elk van de 29 datasets geldt dat de kalibratieset rond de 160 wateren bevat, en de validatie-set rond de 40 wateren.

Deze opzet maakt een goede vergelijking tussen de drie methodes mogelijk. Immers er zijn 29 maal 40 voorspellingen gegenereerd door elk van de drie methodes. Helaas is een vergelijking van methodes niet helemaal eerlijk te maken omdat de neurale netwerken van Haskoning toch gebruik hebben gemaakt van de volledige datasets. Ook hangt de opdeling van 80%/20% per EKR samen met de keuze van hun modellen. Die keuze is zo genomen dat de **moeilijk te voorspellen wateren** terecht zijn gekomen in de validatie-set. Anders gezegd, de verdeling 80% / 20% is niet random gekozen.

Op zich is dat geen probleem voor de PUNNs van Witteveen+Bos of voor de regressiebomen van PBL. Het betekent alleen dat de voorspel-performance in werkelijkheid **beter** zal zijn dan hier geanalyseerd (zie de tabellen in Appendix A). Voor Haskoning zou dat effect nog sterker kunnen zijn, maar analyse van een aantal nieuwe modellen, berekend op een geheel random gekozen 80%/20% verdeling liet een vergelijkbare voorspel-performance zien. Besloten is om de aanvankelijke voorspelperformance van Haskoning mee te nemen in de vergelijking van methodes. Helaas, blijft het zo dat enige voorzichtigheid geboden blijft.

Een samenvatting van de huidige situatie is als volgt:

- de regressiebomen zijn geschat op de kalibratie-set. Data uit de validatie-set zijn hierbij niet gebruikt. De voorspellingen in de validatie-set zijn daarmee zuivere voorspellingen. Zie eerste tabel in Appendix A voor de resultaten op 29 EKR-deelmaatlatten. Regressiebomen zijn ook geschat op **alle** data (dus geschat op kalibratie-set plus validatie-set). Deze laatste set van 29 bomen zijn gestuurd naar Deltares (Erwin Meijer) om in de KRW-verkenner te hangen.
- voor de training van de neurale netwerken van Royal Haskoning is uit de gehele dataset die 80% van de data als kalibratieset gebruikt die de beste neurale netwerken leverden. Data uit de validatie-set zijn hierbij dus ook gebruikt. De voorspellingen in de validatie-set zijn daarmee geen zuivere voorspellingen. Zie tweede tabel in Appendix A voor de resultaten op 29 EKR-deelmaatlatten. Deze neurale netwerken (dus geschat op kalibratie-set plus validatie-set) zijn gestuurd

naar Deltares om in de KRW-verkenner te hangen. Omdat de keuze van kalibratie-set (80%) en validatie-set (20%) wellicht in het nadeel zijn van de neurale netwerken, is voor een gedeelte van de 29 EKR's een random verdeling van 80%/20% gemaakt, waarbij de modellen zijn geschat op de kalibratie-set, zonder de validatie-set te gebruiken. Resultaten bleken vergelijkbaar aan die volgens de eerste opzet.

- de neurale netwerken van Witteveen+Bos zijn geschat op de kalibratie-set. Data uit de validatie-set zijn hierbij niet gebruikt. De voorspellingen in de validatie-set zijn daarmee zuivere voorspellingen. Zie derde tabel in Appendix A voor de resultaten op 29 EKR-deelmaatlaten. Deze neurale netwerken (dus geschat op **alleen** de kalibratie-set) zijn gestuurd naar Deltares om in de KRW-verkenner te hangen.

2.2 Resultaten

De resultaten van de voorspel-exercitie zijn samengevat in de tabellen in Appendix A. De voorspel-performance is geanalyseerd door te kijken naar drie grootheden: de Coefficient of Determination (CoD), de Root Mean Squared Error (RMSE) en het percentage voorspellingen dat ligt binnen een range van ± 0.10 van de werkelijke EKR-maatlat-waarde.

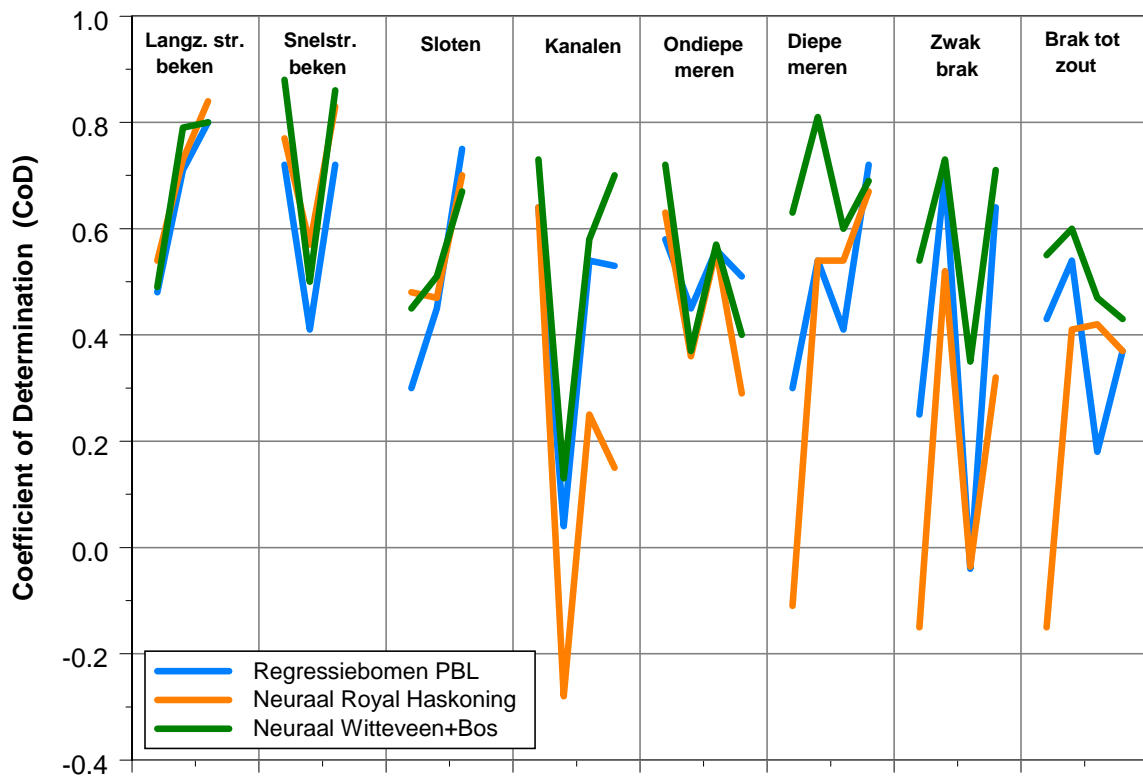
De CoD geeft aan hoe goed het model voorspelt ten opzichte van een methode die voor elke voorspelling het gemiddelde neemt van de validatie-set. Bij een CoD-waarde van 0.0 is de voorspelling even goed als die constante. Bij een CoD van 1.0 is de voorspelling van het model in kwestie perfect. Voor elk van de drie methodes (regressiebomen en twee neurale netwerken) zijn 29 CoD's bepaald. De waarden zijn grafisch weergegeven in **Figuur 1**.

Enkele opvallende zaken uit de figuur:

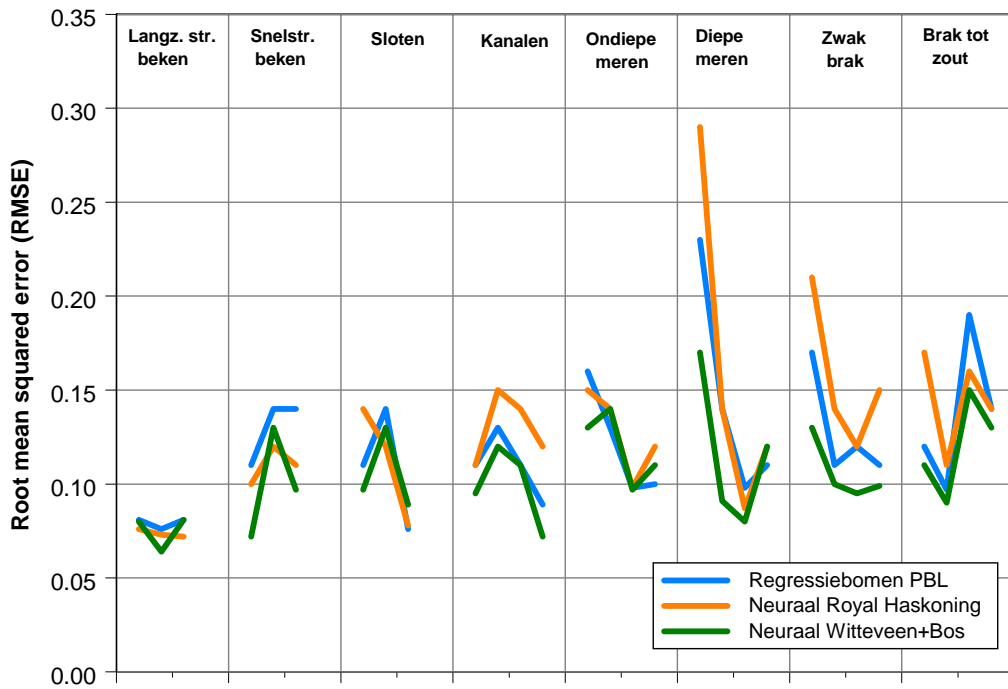
- Neurale netwerken van Witteveen+Bos komen er het beste uit. Dat geldt vooral voor de laatste drie watertypen: diepe meren, zwak-brakke wateren, en brak tot zout wateren.
- De data voor beken blijken het best te voorspellen van alle watertypen (CoD's van 0.50 tot 0.90).
- De neurale netwerken van Haskoning vertonen vijf CoD's die onder de 0.0 duiken. Regressiebomen duiken eenmaal onder de 0.0. Witteveen+Bos geen enkele maal.
- Er is een EKR-deelmaatlat die alle drie methoden slecht voorspellen, namelijk **waterflora in kanalen**.

Een tweede kengetal is het 'percentage binnen 0.10'. Dit kengetal is anders van karakter dan de CoD, in die zin dat hij ongevoelig is voor incidenteel zeer slechte voorspellingen.

Immers zit een voorspelling buiten die 0.10-range, dan maakt het niet meer uit *hoeveel* hij daar buiten viel. Met andere woorden, dit kengetal is ongevoelig voor uitbijters in de voorspellingen. De resultaten liggen wat dichter bijeen dan voor CoD. Zie bovenste paneel in **Figuur 2**. Desalniettemin is voor drie watertypen de PUNN-benadering duidelijk de beste: kanalen, diepe meren en zwak-brakke wateren.



Figuur 1 CoD-waarden voor drie methodes en 29 EKR-deelmaatlatten. De volgorde van de 29 EKR's op de x-as is die zoals weergegeven in de tabellen van Appendix A.



Figuur 2 Voorspel-performance voor 29 EKR's volgens het 'percentage binnen 0.10' (boven) en de RMSE (onder). De volgorde van 29 EKR's op de x-as is gelijk aan die in de tabellen in Appendix A.

Het onderste paneel in figuur 2 geeft de voorspel-resultaten voor de gemiddelde voorspelfout (RMSE). Deze figuur hangt nauw samen met die voor de CoD, maar dan gespiegeld (zie definitie van CoD in Appendix A). Resultaten zijn daarom vergelijkbaar met die in Figuur 1.

De resultaten uit de figuren 1 en 2 zijn samen gevat in **Tabel 1**. In de tabel zijn per methode het gemiddelde van alle 29 CoD's berekend, idem voor RMSE en idem voor 'percentage binnen 0.10'. De tabel laat zien dat de indruk die figuren 1 en 2 geven, ook getalsmatig kloppen. De PUNNs van Witteveen+Bos komen er voor elk van drie kengetallen als beste uit.

Met de *gepaarde toets van Wilcoxon* is getest of de verschillen in tabel 1 ook statistisch significant zijn. Het blijkt dat voor elk van drie voorspel-indicatoren (perc. binnen 0.10, RMSE en CoD) de resultaten van Witteveen+Bos significant hoger scoren dan zowel PBL en Royal Haskoning. Een vergelijking tussen PBL en Royal Haskoning laat zien dat de cijfers **niet** statistisch significant verschillen. Opmerking: de toegepaste toets van Wilcoxon lijkt op de gepaarde t-toets, maar heeft als voordeel dat geen normale kansverdeling hoeft te worden verondersteld (een 'parameter vrije toets' in jargon).

Samengevat: de voorspelresultaten van Witteveen+Bos zijn significant beter dan die van zowel PBL en Royal Haskoning. De *over-all* resultaten van PBL en Royal Haskoning zijn vergelijkbaar (hoewel verschillend in detail).

Tabel 1 Voorspel-performance gemiddeld over alle 29 EKR-deelmaatlatten.

	Percentage binnen 0.10	RMSE	CoD
PBL	63%	0.121	0.49
Royal Haskoning	64%	0.129	0.41
Witteveen+Bos	68%	0.106	0.60

3 Gebruik van expert judgment bij modellering

Bij twee modellen is er voor gekozen om relaties die door biologen/ecologen onrealistisch worden gevonden, uit te sluiten in de finale modellen. Deze keuze is gemaakt door PBL en door Royal Haskoning. Witteveen+Bos hebben niet een dergelijke eis opgelegd. De belangrijkste onrealistische relaties zijn stijgende EKR-waarden bij stijgende P- of N-concentratie.

Meer in detail heeft Haskoning (en PBL idem) de volgende voorwaarden opgelegd aan hun neurale netwerken. Een netwerk wordt na een trainingscyclus verworpen indien een KRW-maatlat stijgt en één of meer van de volgende situaties geldt

- een toename van de concentratie P, N of BZV of
- een afname van de sinuositeit van meandering of
- een toename van de mate van verstuwing of
- een afname van de mate van beschaduwing of
- het minder natuurlijk worden van de oeverinrichting of
- het minder natuurlijk worden van de peildynamiek of
- een toename van de intensiteit van onderhoud of
- een toename van de scheepvaart.

Het is natuurlijk de vraag of zulke relaties inderdaad onrealistisch en dus ongewenst zijn. Witteveen+Bos stelt zich op het standpunt dat het onder zekere voorwaarden (specifieke combinatie van stuurvariabelen) best mogelijk is dat een EKR-deelmaatlat stijgt bij stijgende P of N. Hoe zeker is die biologische/ecologische judgment?

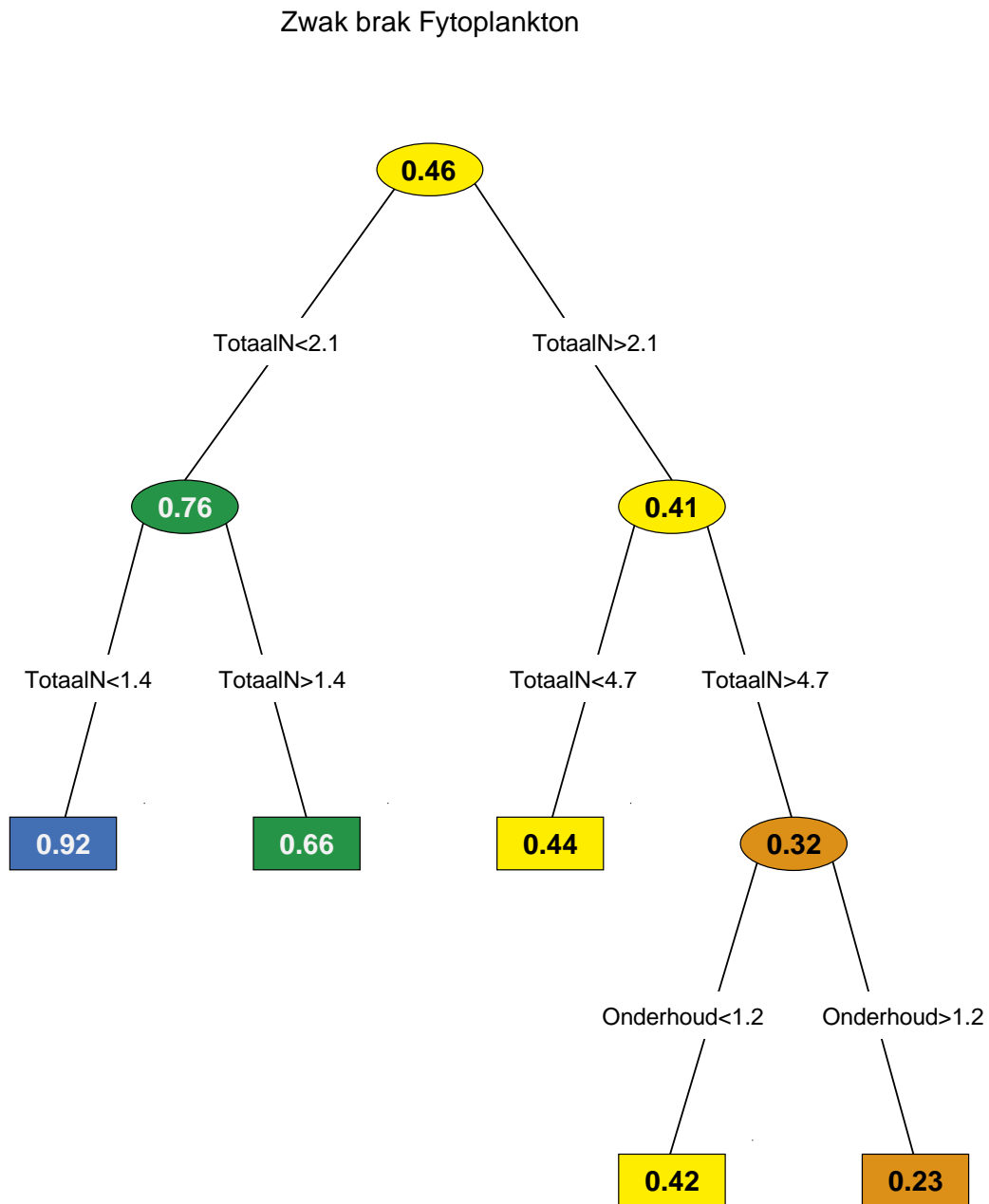
Kortom twee visies op het toelaten van expert judgment, en dus op het wel of niet opleggen van restricties. Voor beiden is wat te zeggen. Het zou in vervolgonderzoek wel interessant zijn om te kijken onder welke omstandigheden die tegen-intuïtieve relaties bij PUNN's optreden. Misschien zijn nieuwe inzichten mogelijk?

Helaas is het niet bekend wat het wel of niet opleggen van restricties voor effect heeft op de *voorspelnauwkeurigheid*. Immers de drie modelleers hebben gekozen voor de ene of de andere aanpak, en niet voor beide benaderingen parallel.

In vervolgonderzoek zal dit punt beter afgestemd moeten worden!

4 Staan modellen een biologische interpretatie toe?

De hier beschreven modellen verschillen in de mate van biologische interpretatie. De helderste interpretatie-mogelijkheden bieden regressiebomen. **Figuur 3** geeft een voorbeeld voor fytoplankton in zwak-brakke wateren.



Figuur 3 Voorbeeld van een regressieboom voor fytoplankton in zwak-brakke wateren.

Als waarden voor de stuurvariabelen bekend zijn, dan kan direct uit de figuur de beste EKR-voorspelling afgelezen worden (de waarden in de ellipsen en uiteindelijk in de rechthoeken). Is bijvoorbeeld in een water de totaal-N-concentratie 1.11 mg/liter, dan is de EKR-voorspelling volgens deze boom 0.92. Is de totaal-N-concentratie hoger dan 4.7 mg/liter, en het onderhoud gelijk aan 1 (= intensief), dan is de EKR-voorspelling 0.42.

De voorspellingen van een regressieboom verlopen schoksgewijs: als een stuurvariabele een drempelwaarde over- of onderschrijdt, zal de voorspelling veranderen. Kleine wijzigingen in stuurvariabelen kunnen dus leiden tot dezelfde voorspelling. Verklaring voor deze schoksgewijze voorspellingen is de onzekerheid in een voorspelling: als de onzekerheid in een voorspelling van, zeg, 0.62 gelijk is aan ± 0.14 , dan levert de aanpassing van 0.62 aan een kleine wijziging in een stuurvariabele niet veel voorspelvoordeel op.

Ook PUNNs laten een interpretatie toe, zij het lastiger dan regressiebomen. Hierna geven we twee voorbeelden, voor fytoplankton. De formules zijn tamelijk ingewikkeld en laten simpele relaties niet direct zien. Wel is direct duidelijk welke stuurvariabelen een rol spelen. Maar of in de bovenste formule de EKR kan stijgen bij stijgende totaal-N is niet eenvoudig te zien (totaal N komt driemaal voor in de formule, met steeds andere wegingen).

Interessant is hier dat de regressieboom uit figuur 3 vergeleken kan worden met de onderste formule in **figuur 4**.

$$EKR = -0.735 + 0.996 \frac{1}{N^{0.288}} + 0.0029N^{1.54} + 0.0992 \frac{N^{0.311}}{P^{0.409}}$$

$$EKR = -0.690 - 0.0015 \frac{O^{3.26}}{P} - 2973 \frac{C^{3.48} P^{0.435} N^{0.692}}{Pe^{4.71} Cl^{1.47}} + 1.51 \frac{1}{Pe^{0.0574} P^{0.0887} N^{0.154}} + 2481 \frac{C^{3.50} P^{0.451} N^{0.697}}{Pe^{4.24} Cl^{1.46}}$$

Figuur 4 Wiskundige vergelijkingen voor twee PUNNs: boven voor fytoplankton in diepe meren, onder voor fytoplankton in zwak-brakke wateren. Codering : N= totaal N, P= totaal P, O= onderhoud, C= connectiviteit, Pe= Peildynamiek, Cl= Chloride.

De regressieboom gebruikt alle de stuurvariabelen totaal N en Onderhoud. De PUNN gebruikt deze variabelen ook, maar gebruikt daarnaast nog totaal P, Chloride, Peildynamiek en Connectiviteit (geleidbaarheid). Voor dit voorbeeld (fytoplankton in zwak-brakke wateren) is de voorspel-performance van de PUNN duidelijk beter (zie tabellen in Appendix A).

De neurale netwerken van Royal Haskoning kunnen ook in wiskundige vergelijkingen worden samengevat, zij het veel ingewikkelder dan getoond in figuur 4. In de praktijk worden deze formules niet geanalyseerd. Het gevolg is dat deze neurale netwerken daarom geen biologisch/ecologisch inzicht verschaffen.

5 Zijn alle wateren/EKR's even goed te voorspellen?

De figuren 1 en 2 laten zien welke wateren en/of EKR's goed dan wel slecht te voorspellen zijn. Wat CoD betreft zou je kunnen stellen dat modellen die slechter zijn dan CoD-waarden onder de 0.20, niet veel waarde hebben. Immers ze voorspellen amper beter dan de voorspeller die altijd een constante invult (het gemiddelde van alle wateren in de validatieperiode). De vraag is natuurlijk wat een goede ondergrens is voor CoD. Ik denk dat dat een kwestie van smaak is. De statistiek kan hier ook geen eenduidig antwoord op geven.

Ik stel voor om een waarde van 0.20 te hanteren. We zien dat regressiebomen en PUNNs hier eenmaal onderduiken bij kanalen. De neurale netwerken van Haskoning duiken er vijf maal onder.

Mijn voorstel zou zijn om in de KRW-verkenner 28 EKR's te laten voorspellen met PUNNs, 26 EKR's met regressiebomen en 23 EKR's met de neurale netwerken van Royal Haskoning. Zie figuur 1 welke EKR-deelmaatlaten weg zouden moeten vallen.

6 Rol van ingeschatte EKR-deelmaatlat-waarden

Niet alle EKR-waarden in de 29 datasets bestaan uit metingen. Een deel is ingeschat omdat er te weinig waarnemingen aanwezig waren. Een beoordeling van de mate van inschatten is gegeven in de tabellen van Appendix A (zie laatste kolom). Het blijkt dat 13 van de 29 datasets leiden aan het euvel van inschattingen.

Een interessante vraag is of het wel of niet aanwezig zijn van een groot aantal inschattingen de voorspel-performance beïnvloedt. Om dat na te gaan zijn de resultaten uit tabel 1 opnieuw berekend voor de deelverzameling met veel inschattingen (13 EKR-maatlatten) en weinig inschattingen (16 EKR-deelmaatlatten). Het resultaat is getoond in **tabel 2**.

Tabel 2 Voorspel-performance gemiddeld over twee delverzamelingen uit de 29 EKR-deelmaatlatten: EKR-deelmaatlatten met weinig inschattingen (16 EKR's) en met veel inschattingen (13 EKR's).

	Percentage binnen 0.10		RMSE		CoD	
	weinig ingeschat	veel ingeschat	weinig ingeschat	veel ingeschat	weinig ingeschat	veel ingeschat
PBL	59%	68%	0.130	0.110	0.40	0.61
Royal Haskoning	60%	70%	0.139	0.118	0.32	0.52
Witteveen+Bos	63%	74%	0.113	0.097	0.54	0.67

De tabel laat zien dat de voorspelperformance duidelijk verbetert wanneer veel data zijn ingeschat. Dat geldt voor alle drie de voorspel-indicatoren en voor elk van de drie methodes. Dit is ook logisch omdat de ingeschatte waarden bepaald zijn met simpele regels 'in het achterhoofd'. Deze regels zijn dan ook weer eenvoudig terug te schatten door de drie methodes.

7 Samenvatting en implementatie in de KRW-verkenner

7.1 Samenvatting

De resultaten van deze notitie kunnen als volgt worden samengevat:

- De voorspel-performance is samengevat in tabel 1. Het blijkt dat voor alle criteria de PUNNs van Witteveen+Bos het beste voorspellen. Regressiebomen en neurale netwerken van Royal Haskoning eindigen gelijk op de tweede plaats. Genoemde verschillen zijn statistisch significant.
- De modelbenaderingen verschillen in het stellen van restricties aan de relaties tussen de EKR-deelmaatlat enerzijds, en de bijbehorende stuurvariabelen anderzijds. Royal Haskoning en PBL hebben restricties opgelegd aan hun modellen. Zo kan het bij hun niet zo zijn dat een EKR-waarde stijgt bij een stijgende totaal-N-concentratie. Bij de PUNN-benadering van Witteveen+Bos zijn zulke restricties in het geheel niet opgelegd. Helaas weten we niet hoe de voorspel-performance beïnvloed wordt door het wel of niet meenemen van deze restricties.
- De biologische interpretatie van regressiebomen is het eenvoudigst (figuur 3), gevolgd door de PUNN-formules van Witteveen+Bos (figuur 4). De neurale netwerken van Royal Haskoning kunnen weliswaar samengevat worden in wiskundige vergelijkingen, maar die zijn zeer ingewikkeld en laten geen eenvoudige interpretatie toe.
- De voorspelkarakteristiek van de methodes verschilt. Bij de neurale netwerken verandert de EKR-deelmaatlat-waarde continu bij het veranderen van stuurvariabele-waarden. Dat geldt ook bij geringe veranderingen. Bij regressiebomen verspringt de voorspelling discontinu wanneer de stuurvariabelen bepaalde drempelwaarden over- of onderschrijden. Bij kleine veranderingen in stuurvariabelen kan de voorspelling dus gelijk blijven bij regressiebomen.

Bovengenoemde punten zijn samengevat in tabel 3.

Tabel 3 Samenvatting van vier eigenschappen van de onderzochte modelbenaderingen.

	Voorspel-performance CoD	Expert judgment gebruikt?	Biologisch interpreteerbaar?	Continue relatie EKR en stuurvariabelen?
PBL	0.49	ja	ja	nee
Royal Haskoning	0.41	ja	nee	ja
Witteveen+Bos	0.60	nee	ja	ja

Tabel 3 laat zien dat de beoordeling van de verschillende modellen afhangt van welke eigenschappen men belangrijk vindt. Als men het opleggen van restricties niet belangrijk vindt, of zelfs ongewenst, dan zijn de PUNN's de beste keuze. Vindt men deze restricties juist wel belangrijk, of zelfs noodzakelijk, dan komen regressiebomen en de neurale netwerken van Royal Haskoning meer in aanmerking, ondanks een mindere voorspel-performance.

Helaas is dit punt van het incorporeren van restricties op basis van biologische/ecologische inzichten (expert judgment) niet helder besproken bij de start van dit project. We kunnen nu ook helaas niet zeggen of en hoe restricties de voorspel-performance nadelig beïnvloeden. In toekomstig werk moeten hier goede afspraken over gemaakt worden.

7.2 Implementatie KRW-verkenner

Hoe zouden de drie methodes in de KRW-verkenner geïmplementeerd kunnen worden? Er zijn vier mogelijkheden:

- Kies de beste methode uit tabel 2 en hang die in de KRW-verkenner.
- Hang alle drie methoden in de KRW-verkenner en zet de beste methode als default. Gebruikers kunnen ook de voorspellingen bekijken van de andere twee methoden.
- Idem, maar alleen geselecteerde gebruikers kunnen de voorspellingen bekijken van de andere twee methoden.
- Kies geen beste methode uit tabel 3, maar bereken het gemiddelde van de drie voorspellingen. De software geeft dan dit gemiddelde met een ondergrens (de laagste voorspelling) en een bovengrens (de hoogste voorspelling). Dit zou men de "IPCC-methode" kunnen noemen. De kracht van drie modellen wordt expliciet gebruikt.

Er bestaat nog een meningsverschil over de wijze van implementatie van de modellen die in de KRW/verkenner worden gehangen. Moeten ze geschat worden over **alle** data, of alleen over de validatieperiode? Vanuit statistisch oogpunt is een groter aantal wateren altijd beter. Schattingen worden nauwkeuriger bij een toenemende N (hier het aantal wateren per EKR-deelmaatlat).

Het PBL heeft daarom een aparte set regressiebomen geleverd die geschat zijn over alle data (dus validatieperiode plus kalibratieperiode). Witteveen+Bos hebben dat niet gedaan. De getrainde modellen van Royal Haskoning kunnen beschouwd worden als representatief voor alle data. Hierover bestaat dus een verschil van inzicht.

Ook dit is een punt dat bij aanvang van het project niet bediscussieerd is. Het is aan te bevelen om bij vervolgonderzoek hier een duidelijk standpunt over in te nemen.

7.3 Tot slot

Tot slot nog twee opmerkingen. Niet alle modellen voorspellen goed. Voorgesteld wordt om modellen die een CoD-waarde hebben slechter dan 0.20, weg te laten (of te blokkeren) in de KRW-verkenner. Zie figuur 1 voor EKR-deelmaatlaten waarvoor deze opmerking geldt.

Een tweede opmerking volgt uit §6. Bij een aantal EKR-deelmaatlaten zijn EKR-waarden van wateren *ingeschat* in plaats van *gemeten*. Het percentage inschattingen is weliswaar verbeterd ten opzichte van de studie uit 2009, maar toch nog steeds redelijk hoog: bij 13 van de 29 EKR-deelmaatlaten speelt het een rol. Gevonden is dat de drie methodes duidelijk beter voorspellen voor EKR's met veel inschattingen (tabel 2). Het is dus mogelijk dat wanneer in de toekomst *alle* datasets bestaan uit metingen (dus geen ingeschatte data meer), de voorspel-performance er **minder goed** uit ziet dan getoond in de figuren 1 en 2.

Daarentegen is de voorspel-performance zoals in detail getoond in de tabellen van Appendix A juist weer een *onderschatting* van de werkelijke performance. Dat komt door de niet-*random* keuze van de 29 validatiesets. Juist moeilijk te voorspellen wateren zijn in de validatiesets terecht gekomen. In de toekomst moet de opdeling "80% / 20%" geheel *random* worden uitgevoerd! De voorspellingen zullen dan **beter** worden. Maar hoeveel, is lastig te kwantificeren.

Informatie uit deze notitie mag worden overgenomen op voorwaarde van bronvermelding: "PBL-notitie De ecologische kwaliteit van Nederlands oppervlaktewater: een vergelijking van drie modelleringstechnieken, rapportnummer 500208002, Den Haag: Planbureau voor de Leefomgeving."

Appendix A Voorspellingen voor de validatieperiode in detail

De definities van de kengetallen R^2 , perc. binnen 0.10, RMSE en coef of determ. (afgekort als CoD) zijn als volgt:

1. ' R^2 ' is de gebruikelijke correlatie-coëfficiënt, gekwadrateerd. Deze geeft aan hoe goed het **lineaire** verband is tussen X (bij ons de metingen) en Y (bij ons de voorspellingen, door gekozen model). Een waarde van 1.0 geeft aan dat het verband perfect lineair is, een waarde van nul dat er helemaal geen lineair verband is.
2. 'Perc. binnen 0.10' staat voor het percentage van alle data N waarvoor geldt dat de gemeten EKR minder dan 0.10 afwijkt van de voorspelde EKR. Merk op dat dit percentage ongevoelig is voor incidentele (zeer) slechte voorspellingen.
3. 'RMSE' is de Root Mean Squared Error. Deze is als volgt gedefinieerd:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$$

In de formule staat y_i voor de gemeten EKR in water i . Bij een perfecte voorspelling geldt dat de RMSE-waarde gelijk is aan nul. RMSE staat voor de gemiddelde voorspelfout. Merk op dat hele slechte voorspellingen door het kwadraat in de formule extra zwaar mee wegen, dit in tegenstelling tot het percentage onder punt 2.

4. 'Coef. of determ.' (CoD) staat voor de *coefficient of determination*. Deze kan eenvoudig uitgerekend worden vanuit de RMSE. De formule voor de CoD is

$$CoD = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} = 1 - \frac{RMSE^2}{\text{var}(y_i)} \cdot \frac{N}{N-1}$$

met ' $\text{var}(y_i)$ ' de variantie van de metingen in de gekozen dataset, ' \hat{y}_i ' de voorspellingen met neurale netwerken of regressiebomen, en ' \bar{y} ' het gemiddelde van de metingen. De CoD lijkt veel op R^2 onder punt 1 maar wijkt in de validatieperiode toch vaak af van R^2 . Dat komt omdat de CoD kijkt of voorspellingen en metingen precies liggen op de lijn van (0,0) en (1,1): de een-op-een-lijn.

Tabel A.1 Voorspelnaauwkeurigheid regressiebomen (PBL).

EKR	Trainingsset				Validatieset				Veel ingeschat ?
	R ²	perc. binnen 0.10	RMSE	coef of determ	R ²	perc. binnen 0.10	RMSE	coef of determ	
1 Langzaam stromende beken									
Waterflora	0.76	81%	0.074	0.76	0.49	75%	0.081	0.48	nee
Macrofauna	0.87	90%	0.064	0.87	0.73	88%	0.076	0.71	nee
Vissen	0.85	88%	0.065	0.85	0.81	85%	0.081	0.80	de helft
2 Snelstromende beken									
Waterflora	0.90	85%	0.065	0.90	0.73	67%	0.11	0.72	de helft
Macrofauna	0.74	67%	0.11	0.72	0.47	62%	0.14	0.41	nee
Vissen	0.76	70%	0.11	0.75	0.72	69%	0.14	0.72	grotendeels
3 Sloten									
Waterflora	0.76	66%	0.10	0.76	0.76	59%	0.11	0.30	gering
Macrofauna	0.75	74%	0.096	0.74	0.48	67%	0.14	0.45	nee
Vissen	0.86	84%	0.064	0.86	0.78	77%	0.076	0.75	grotendeels
4 Kanalen									
Fytoplankton	0.69	74%	0.11	0.71	0.64	63%	0.11	0.64	bijna helft
Waterflora	0.66	65%	0.10	0.72	0.21	45%	0.13	0.04	een derde
Macrofauna	0.79	69%	0.086	0.79	0.58	53%	0.11	0.54	nee
Vissen	0.84	86%	0.065	0.84	0.56	70%	0.089	0.53	grotendeels
5 Ondiepe meren									
Fytoplankton	0.80	64%	0.11	0.80	0.63	59%	0.16	0.58	een derde
Waterflora	0.86	79%	0.079	0.86	0.53	64%	0.13	0.45	grotendeels
Macrofauna	0.74	76%	0.087	0.78	0.53	67%	0.10	0.56	vrijwel niet
Vissen	0.81	77%	0.090	0.84	0.50	72%	0.10	0.51	grotendeels
6 Diepe meren									
Fytoplankton	0.79	62%	0.13	0.80	0.51	35%	0.23	0.30	vrijwel niet
Waterflora	0.87	76%	0.082	0.87	0.58	54%	0.14	0.54	ruim helft
Macrofauna	0.86	88%	0.066	0.86	0.50	62%	0.098	0.41	vrijwel niet
Vissen	0.85	83%	0.072	0.85	0.76	69%	0.11	0.72	vrijwel niet
7 Zwak brakke wateren									
Fytoplankton	0.45	52%	0.16	0.45	0.25	41%	0.17	0.25	vrijwel niet
Waterflora	0.59	62%	0.12	0.60	0.74	59%	0.11	0.70	de helft
Macrofauna	0.69	79%	0.098	0.72	0.26	59%	0.12	-0.04	vrijwel niet
Vissen	0.54	68%	0.11	0.65	0.65	63%	0.11	0.64	grotendeels
8 Brak tot zoute wateren									
Fytoplankton	0.59	61%	0.12	0.61	0.47	59%	0.12	0.43	vrijwel niet
Waterflora	0.74	81%	0.080	0.74	0.58	76%	0.097	0.54	grotendeels
Macrofauna	0.74	65%	0.11	0.74	0.29	46%	0.19	0.18	niet
Vissen	0.58	72%	0.097	0.58	0.36	59%	0.14	0.37	grotendeels

Tabel A.2 Voorspelnaauwkeurigheid Neurale Netwerken (Royal Haskoning/DHV).

EKR	Trainingsset				Validatieset				Veel ingeschat ?
	R ²	perc. binnen 0.10	RMSE	coef of determ	R ²	perc. binnen 0.10	RMSE	coef of determ	
1 Langzaam stromende beken									
Waterflora	0.81	88%	0.066	0.81	0.62	78%	0.076	0.54	nee
Macrofauna	0.88	91%	0.06	0.88	0.78	88%	0.073	0.73	nee
Vissen	0.87	90%	0.061	0.87	0.85	88%	0.072	0.84	de helft
2 Snelstromende beken									
Waterflora	0.91	87%	0.063	0.91	0.8	79%	0.10	0.77	de helft
Macrofauna	0.75	68%	0.10	0.77	0.68	64%	0.12	0.57	nee
Vissen	0.82	72%	0.093	0.82	0.84	67%	0.11	0.83	grotendeels
3 Sloten									
Waterflora	0.79	73%	0.089	0.81	0.51	44%	0.14	0.48	gering
Macrofauna	0.77	69%	0.093	0.76	0.50	59%	0.12	0.47	nee
Vissen	0.89	93%	0.057	0.89	0.73	77%	0.078	0.70	grotendeels
4 Kanalen									
Fytoplankton	0.86	82%	0.076	0.86	0.69	73%	0.11	0.64	bijna helft
Waterflora	0.68	67%	0.099	0.67	0.19	55%	0.15	-0.28	een derde
Macrofauna	0.78	72%	0.090	0.78	0.41	53%	0.14	0.25	nee
Vissen	0.81	86%	0.070	0.81	0.43	73%	0.12	0.15	grotendeels
5 Ondiepe meren									
Fytoplankton	0.78	61%	0.12	0.77	0.69	64%	0.15	0.63	een derde
Waterflora	0.84	79%	0.086	0.84	0.45	62%	0.14	0.36	grotendeels
Macrofauna	0.73	72%	0.089	0.73	0.56	72%	0.098	0.56	vrijwel niet
Vissen	0.83	75%	0.086	0.83	0.46	69%	0.12	0.29	grotendeels
6 Diepe meren									
Fytoplankton	0.66	54%	0.16	0.65	0.18	42%	0.29	-0.11	vrijwel niet
Waterflora	0.87	75%	0.081	0.87	0.61	65%	0.14	0.54	ruim helft
Macrofauna	0.86	89%	0.068	0.86	0.54	77%	0.087	0.54	vrijwel niet
Vissen	0.85	85%	0.071	0.85	0.71	65%	0.12	0.67	vrijwel niet
7 Zwak brakke wateren									
Fytoplankton	0.70	72%	0.12	0.69	0.27	32%	0.21	-0.15	vrijwel niet
Waterflora	0.81	81%	0.084	0.80	0.62	59%	0.14	0.52	de helft
Macrofauna	0.66	75%	0.10	0.69	0.32	66%	0.12	-0.036	vrijwel niet
Vissen	0.71	77%	0.091	0.69	0.46	71%	0.15	0.32	grotendeels
8 Brak tot zoute wateren									
Fytoplankton	0.74	67%	0.10	0.73	0.27	46%	0.17	-0.15	vrijwel niet
Waterflora	0.84	88%	0.063	0.84	0.53	70%	0.11	0.41	grotendeels
Macrofauna	0.68	58%	0.12	0.70	0.44	49%	0.16	0.42	niet
Vissen	0.68	83%	0.086	0.67	0.38	57%	0.14	0.37	grotendeels

Tabel A.3 Voorspelnaauwkeurigheid PUNNs (Witteveen+Bos).

EKR	Trainingsset				Validatieset				Veel ingeschat ?
	R ²	perc. binnen 0.10	RMSE	coef of determ	R ²	perc. binnen 0.10	RMSE	coef of determ	
1 Langzaam stromende beken									
Waterflora	0.81	87%	0.066	0.81	0.54	78%	0.080	0.49	nee
Macrofauna	0.84	85%	0.070	0.84	0.81	85%	0.064	0.79	nee
Vissen	0.86	88%	0.062	0.86	0.84	80%	0.081	0.80	de helft
2 Snelstromende beken									
Waterflora	0.92	94%	0.056	0.92	0.88	85%	0.072	0.88	de helft
Macrofauna	0.77	72%	0.10	0.76	0.52	46%	0.13	0.50	nee
Vissen	0.82	72%	0.093	0.82	0.86	74%	0.097	0.86	grotendeels
3 Sloten									
Waterflora	0.78	66%	0.096	0.78	0.47	69%	0.097	0.45	gering
Macrofauna	0.78	71%	0.088	0.78	0.55	56%	0.13	0.51	nee
Vissen	0.84	88%	0.068	0.84	0.71	79%	0.089	0.67	grotendeels
4 Kanalen									
Fytoplankton	0.81	78%	0.089	0.81	0.77	73%	0.095	0.73	bijna helft
Waterflora	0.72	74%	0.091	0.72	0.32	58%	0.12	0.13	een derde
Macrofauna	0.77	71%	0.091	0.77	0.63	65%	0.11	0.58	nee
Vissen	0.81	85%	0.071	0.81	0.71	80%	0.072	0.70	grotendeels
5 Ondiepe meren									
Fytoplankton	0.85	73%	0.095	0.85	0.79	67%	0.13	0.72	een derde
Waterflora	0.86	75%	0.079	0.86	0.42	59%	0.14	0.37	grotendeels
Macrofauna	0.74	77%	0.087	0.74	0.58	64%	0.097	0.57	vrijwel niet
Vissen	0.85	81%	0.080	0.85	0.44	69%	0.11	0.40	grotendeels
6 Diepe meren									
Fytoplankton	0.69	50%	0.15	0.69	0.64	46%	0.17	0.63	vrijwel niet
Waterflora	0.85	76%	0.086	0.85	0.82	73%	0.091	0.81	ruim helft
Macrofauna	0.86	87%	0.066	0.86	0.61	85%	0.080	0.60	vrijwel niet
Vissen	0.86	90%	0.069	0.86	0.70	69%	0.12	0.69	vrijwel niet
7 Zwak brakke wateren									
Fytoplankton	0.59	54%	0.14	0.59	0.55	51%	0.13	0.54	vrijwel niet
Waterflora	0.81	75%	0.082	0.81	0.76	76%	0.10	0.73	de helft
Macrofauna	0.68	77%	0.10	0.68	0.37	66%	0.095	0.35	vrijwel niet
Vissen	0.77	79%	0.078	0.77	0.71	78%	0.099	0.71	grotendeels
8 Brak tot zoute wateren									
Fytoplankton	0.57	58%	0.12	0.57	0.56	59%	0.11	0.55	vrijwel niet
Waterflora	0.82	84%	0.067	0.82	0.62	73%	0.090	0.60	grotendeels
Macrofauna	0.78	68%	0.10	0.78	0.50	46%	0.15	0.47	niet
Vissen	0.76	86%	0.073	0.76	0.50	57%	0.13	0.43	grotendeels