

Verificatie van publieksverwachtingen

Seijo Kruizinga

Het Nederlandse publiek wordt via vele media en vanuit meerdere bronnen (providers) van weersverwachtingen voorzien. Op internet zijn bijvoorbeeld van tenminste vier providers dagelijks geregionaliseerde verwachtingen, voor de volgende dag, te vinden. Voor de hand liggende vragen met betrekking tot deze verwachtingen zijn: wat is de kwaliteit van deze verwachtingen en is er verschil in kwaliteit tussen deze verwachtingen. Maar je kunt je ook afvragen in hoeverre de regionalisatie op zo'n klein gebied als Nederland gerechtvaardigd is. Uit nieuwsgierigheid ben ik gestart met een project om deze verwachtingen dagelijks vast te leggen en ze in een later stadium te verifiëren. In dit artikel komen de eerste resultaten aan de orde en wordt aandacht besteed aan voorgaande vragen.

Inleiding

Zoals gezegd zijn er op internet van meerdere providers geregionaliseerde verwachtingen beschikbaar. Om bovenstaande vragen te kunnen beantwoorden zul je gedurende een substantiële periode deze verwachtingen en de bijbehorende waarneming moeten registreren. Helaas wordt een gedeelte van deze verwachtingen in de vorm van een plaatje gepubliceerd zodat het niet eenvoudig is om de inhoud voor verdere bewerking vast te leggen. De cijfermatige inhoud moet handmatig worden overgebracht in verificatiebestanden. Bij de start van dit project is dat gedaan door de getalsmatige inhoud dagelijks direct in te voeren in een spreadsheet. Helaas werd al snel duidelijk dat dit mogelijk geen betrouwbare dataset op zou leveren en dat bovendien controle achteraf niet mogelijk was. Vanaf begin november zijn daarom dagelijks om negen uur 's morgens de schermbeelden via een screengrabber opgeslagen zodat de handmatige overdracht in een later stadium kon worden gedaan en bovendien achteraf nog controleerbaar is. Gedeeltelijk bestaan de verwachtingen uit plaatjes van Nederland met daarin voor één of meerdere elementen de verwachting voor de volgende dagen. Andere verwachtingen geven een tabel voor meerdere dagen voor een gegeven locatie.

De schermbeelden die vanaf begin november zijn opgeslagen zijn: de meerdaagse verwachtingen van Weathernews (WNI) voor Eelde en Utrecht, de meerdaagse verwachtingen van Weer Online (WOL) voor Groningen en Utrecht, de (temperatuur)-verwachtingen van Meteo Consult (MC) voor dag 1, 2 en 3, de verwachtingen voor de maximumtemperatuur voor de volgende dag van het KNMI. Deze vier providers werden geselecteerd omdat ze verwachtingen geven voor herkenbare locaties in de omgeving van KNMI meetstations. Voor de waarnemingen werd uitgegaan van de KNMI-pagina met "Opgetreden extremen". In december werd nog de verwachting voor de nachttemperatuur voor de komende nacht van het KNMI toegevoegd.

Alle verwachtingen doen uitspraken over meerdere elementen. Echter alleen de temperatuurverwachtingen zijn bij alle verwachtingen numeriek, en dus verifieerbaar, geformuleerd. Dit verhaal beperkt zich tot de verificatie van de verwachtingen voor de maximumtemperatuur voor de volgende dag. Gezien de hoeveelheid extractiewerk konden slechts twee locaties worden meegenomen in de studie, één in Noord-Oost Nederland (Eelde) de andere locatie in het centrum van Nederland (De Bilt).

De periode waarover hier wordt gerapporteerd loopt van 5 november 2006 tot en met 4 juli 2007. Deze periode werd opgesplitst in twee gedeelten: het koude seizoen en het warme seizoen. Voor de scheidslijn tussen deze twee perioden werd 25 maart 2007 gekozen. Onder meer omdat op die datum de zomertijd inging en toch werd vastgehouden aan 09:00 uur lokaal voor het vastleggen van de verwachtingen. De verwachtingen van voor en na 25 maart komen dus vanuit enigszins verschillende fasen in het productieproces.

In beide perioden ontbreken op meerdere dagen de verwachtingen als gevolg van privé omstandigheden zoals vakantie enzovoorts. Ontbrekende waarnemingen werden aangevuld uit KNMI-tabellen met dagwaarden, die door de Klimatologische Dienst op internet beschikbaar worden gesteld. Weliswaar heeft de maximumtemperatuur in deze tabellen betrekking op de het maximum over de hele dag, 00-24 uur, in plaats van over het tijdvak van 06-18 uur zoals in de extremen, maar in de meeste gevallen zijn die waarden identiek.

Verificatiescores

Om de kwaliteit van temperatuursverwachtingen, over een bepaalde periode, te kwantificeren zijn meerdere kwaliteitskenmerken beschikbaar. De meest bekende zijn de Mean Absolute Error (MAE) en de Root Mean Square Error (RMSE). De MAE is het gemiddelde van de absolute waarde van de verschillen tussen de verwachte waarde en de opgetreden waarde. De RMSE is de wortel uit het gemiddelde van de kwadraten van diezelfde verschillen. Beide kenmerken meten dus in feite de mate waarin de verwachte waarden afwijken van de opgetreden waarden. Naast deze kenmerken worden ook vaak de Bias, het gemiddelde verschil tussen verwacht en opgetreden en de standaarddeviatie (Std) van de verschillen gebruikt. In dit verhaal geven we de voorkeur aan de Bias en de Std omdat deze grootheden zich veel beter lenen voor een verdere statistische analyse. Bovendien kan uit Bias en Std op eenvoudige wijze de RMSE berekend worden volgens:

$$RMSE = \sqrt{(Std^2 + Bias^2)}$$

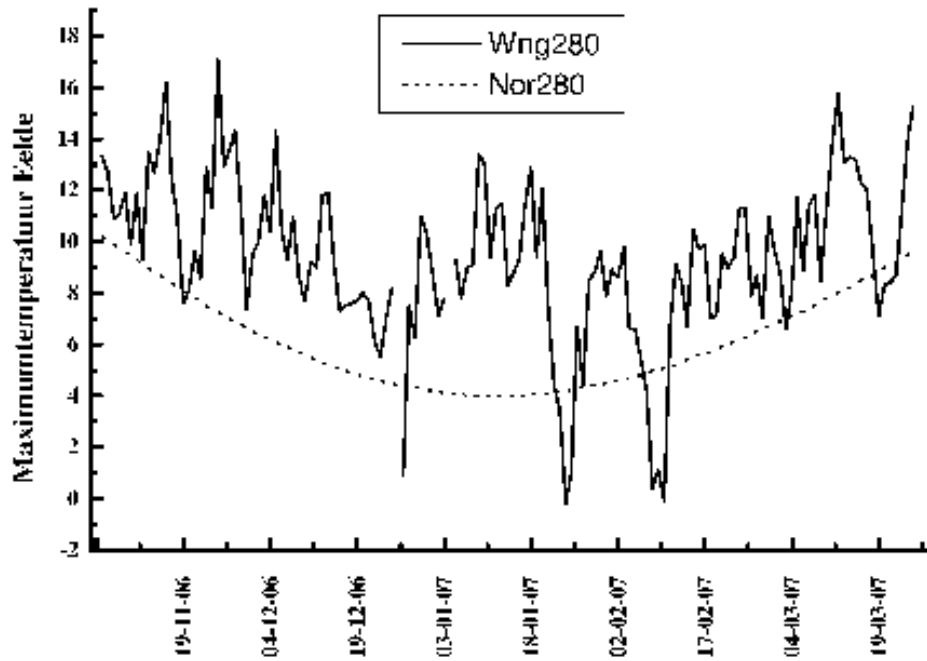
De MAE is ook een combinatie van Bias en STD. Echter de samenhang is niet eenvoudig in een formule uit te drukken.

Het is gebruikelijk om de resultaten voor deze scores te vergelijken met de resultaten die worden behaald met een meestal zeer eenvoudige referentie verwachting. Voor temperatuurverwachtingen blijkt de persistentie (Pers), "de maximumtemperatuur van morgen is gelijk aan die van vandaag", een zeer goede referentie te zijn. De resultaten die hiermee worden behaald zullen dan ook, ter vergelijking, vermeld worden.

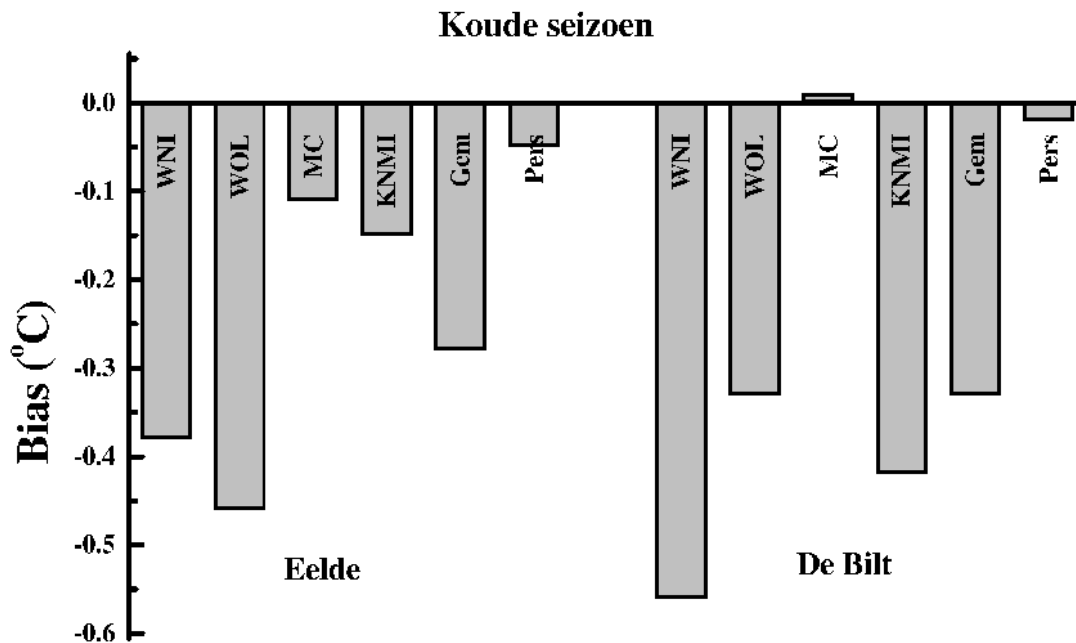
Bij de interpretatie van standaarddeviatie van de fout in de verwachting moeten we nog rekening houden met het feit dat de verwachtingen in hele graden worden geformuleerd. Door deze afronding wordt een extra fout toegevoegd. In de praktijk, bij de huidige kwaliteit van de verwachtingen, blijkt deze extra fout verwaarloosbaar.

Verificatieresultaten voor het koude seizoen

De periode waarover zowel voor De Bilt als voor Eelde gegevens beschikbaar zijn loopt van 5 november 2006 t/m 25 maart 2007. In figuur 1 is het verloop van de maximumtemperatuur in Eelde in beeld gebracht. In deze figuur is tevens het verloop van de normale temperatuur uitgezet. Duidelijk is te zien dat over praktisch de hele periode de maximumtemperatuur in Eelde hoger was dan normaal. Op enkele dagen in deze periode ontbraken één of meer verwachtingen (onderbrekingen in de curve van het temperatuurverloop) zodat uiteindelijk 139 dagen beschikbaar bleven voor de verificatie. Het temperatuurverloop in De Bilt geeft een soortgelijk beeld.



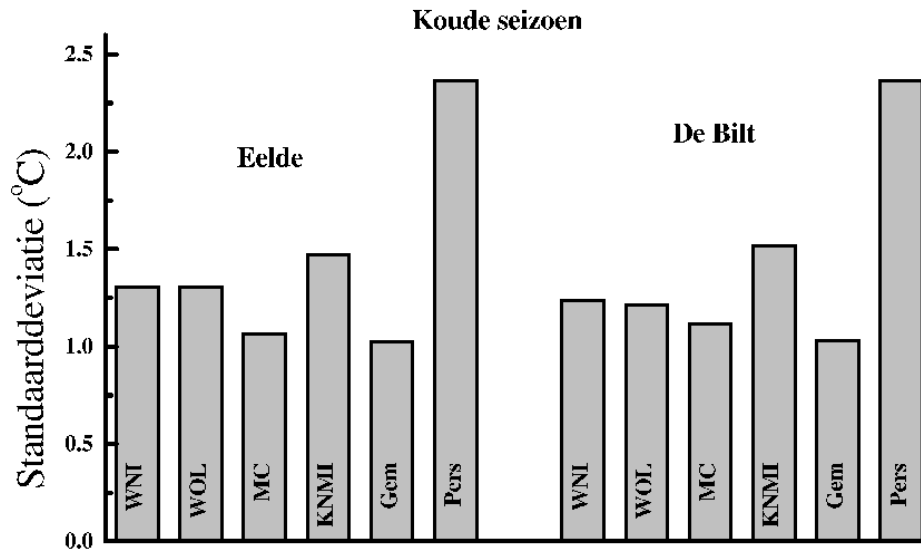
Figuur 1. Verloop van de maximumtemperatuur in Eelde in de periode 5 november 2006 t/m 25 maart 2007.



Figuur 2. Berekende Bias van de verwachtingen voor de maximumtemperatuur voor de locaties Eelde en De Bilt voor de periode van figuur 1.

In figuur 2 zijn de resultaten met betrekking tot de Bias weergegeven voor zowel Eelde als De Bilt. We zien in deze figuur dat de Bias sterk varieert van Provider tot Provider. De Bias van MC is klein voor zowel Eelde als De Bilt. De overige providers vertonen een aanzienlijke

Bias. In deze figuur zijn ook de Bias-gegevens van het gemiddelde van de verwachtingen van de vier providers en van de Persistentie opgenomen. De Bias van de persistentie is zoals verwacht klein. De Bias van het gemiddelde is uiteraard gelijk aan het gemiddelde van de providers. Alhoewel de Bias soms aanzienlijk is zal geen van deze Bias-waarden voor de gebruiker hinderlijk zijn.



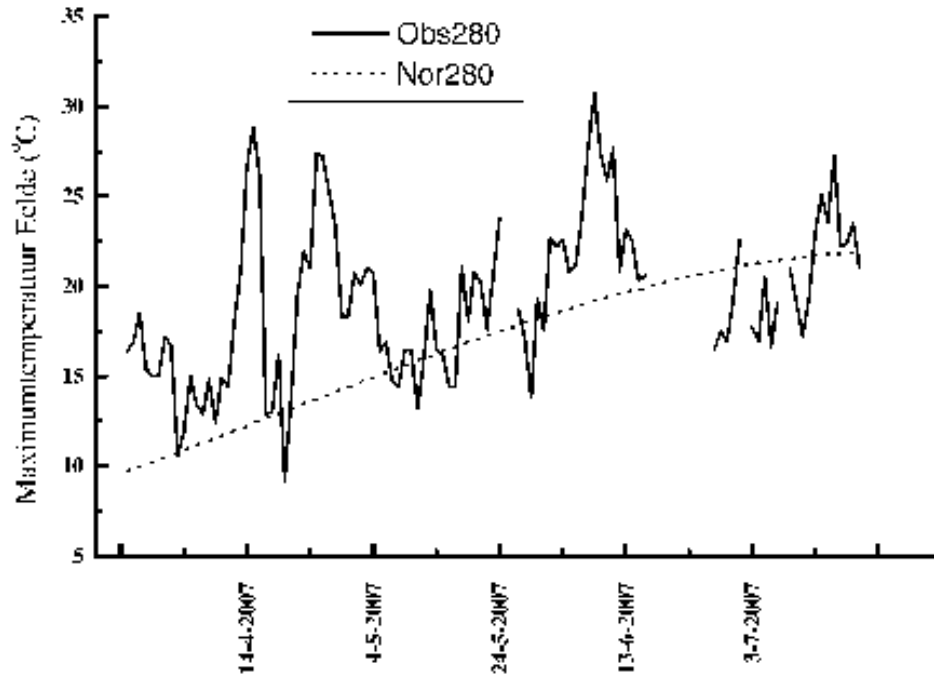
Figuur 3. Berekende standaarddeviaties van de fout in de verwachtingen voor de maximumtemperatuur voor de locaties Eelde en De Bilt voor de periode uit figuur 1.

In figuur 3 is de standaarddeviatie (Std) van de fout in de verwachtingen, over dezelfde periode, voor de vier providers en eveneens voor de gemiddelde verwachting en voor de Persistentie geplote. Allereerst valt op dat alle verwachtingen duidelijk beter scoren (een kleinere Std hebben) dan de Persistentie. Bedenk daarbij ook nog dat de maximumtemperatuur van vandaag die we gebruiken voor de verwachting voor morgen nog niet bekend is als de verwachtingen van de providers worden uitgebracht. Verder toont de figuur voor beide locaties hetzelfde beeld. De Std van MC is het laagst, die van het KNMI het hoogst en WOL en WNI zitten daar ongeveer midden tussen. De Std van de gemiddelde verwachting is marginaal lager dan de laagste provider. Voor de interpretatie van de standaarddeviaties is het nuttig om je te realiseren dat een standaarddeviatie van 1.04 °C (MC, Eelde) betekent dat de fout in de verwachting makkelijk kan variëren +2 graden tot -2 graden, dat is dus aanzienlijk meer dan de Bias-waarden uit figuur 2. In figuur 3 valt op dat er een aanzienlijk verschil is tussen de standaarddeviaties van de verwachtingen van de vier providers. Dat betekent niet zonder meer dat de onderliggende kwaliteit van de verwachtingen ook verschillend zijn. De standaarddeviaties worden hier bepaald op basis van een reeks van 139 waarden. Ook als de onderliggende kwaliteit van alle providers hetzelfde is en dus de standaarddeviaties gelijk zouden moeten zijn kan als gevolg van het toeval nog een verschil optreden. We zullen hier nader op ingaan in de slotparagraaf.

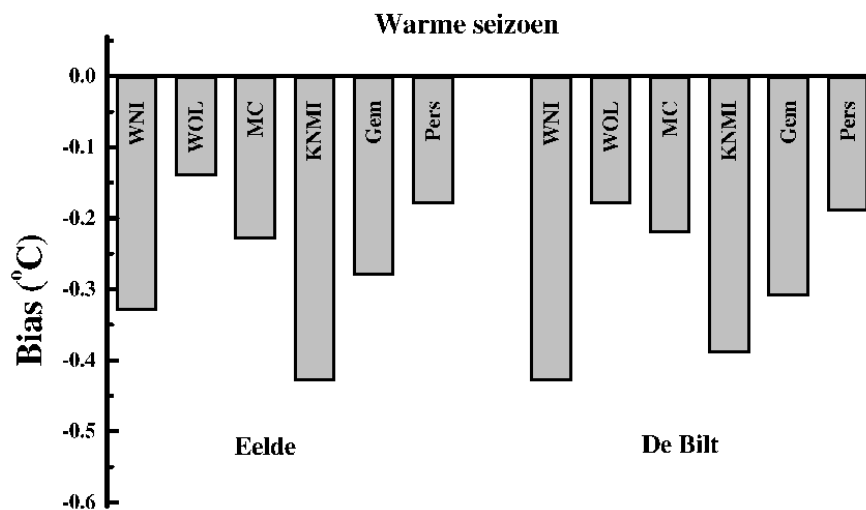
Verificatieresultaten voor het warme seizoen.

De reeks van gegevens strekt zich nu uit van 26 maart t/m 20 juli 2007. Figuur 4 geeft een beeld van het verloop van de maximumtemperatuur in Eelde. Wederom zijn de waargenomen maximumtemperaturen overwegend hoger dan het klimatologisch gemiddelde. Er zijn in deze periode meer en aanzienlijk langere onderbrekingen, vooral ten gevolge van vakantie. In totaal resteerden nog 103 dagen waarover een verificatie kon worden uitgevoerd.

In figuur 5 vinden we de resultaten voor de Bias. De resultaten variëren van $-0,2$ °C tot omstreeks $-0,4$ °C. WNI en KNMI hebben de grootste Bias zowel in Eelde als in De Bilt. Opvallend is de Bias van de Persistentie ($\sim -0,15$ °C). Dit wordt waarschijnlijk veroorzaakt doordat in het grootste deel van de periode het klimatologisch gemiddelde sterk stijgt. In figuur 6 zijn de resultaten voor de Std uitgezet. Alle standaarddeviaties zijn duidelijk hoger dan de overeenkomstige gegevens in de koude periode. De Std's van de providers zijn relatief ongeveer evenveel gestegen als de Std van de Persistentie. De verschillen tussen de providers zijn aanzienlijk kleiner.



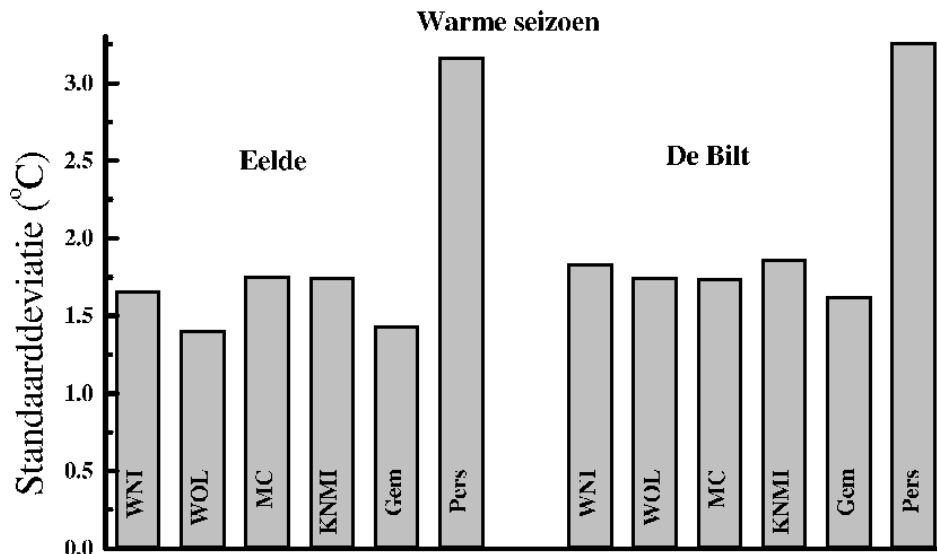
Figuur 4. Als figuur 1 maar dan voor de periode 26 maart 2007 t/m 20 juli 2007.



Figuur 5. Als figuur 2 maar dan voor de periode van figuur 4.

Zijn regionale verwachtingen zinvol?

Alle vier de providers geven temperatuurverwachtingen voor meerdere locaties in Nederland. Soms zelfs tot op het niveau van de postcode. Om een indruk te krijgen in hoeverre die regionalisatie zinvol is zijn de verschillen in verwachte temperaturen voor De Bilt en Eelde vergeleken met de opgetreden verschillen. Voor de opgetreden verschillen vinden we in de koude periode een gemiddeld verschil van $-0,25$ °C en de standaarddeviatie van de opgetreden verschillen bedraagt $1,20$ °C. Voor de warme periode vonden we $-0,77$ °C en $1,58$ °C. De standaarddeviaties zijn vergeleken met de standaarddeviaties van de verwachtingen betrekkelijk klein. Bij de analyse op dagbasis is daarom alleen gekeken naar de gegevens van de beste provider, in de koude periode is dat Meteo Consult en in de warme periode is dat Weer Online. In de figuren 7 en 8 zijn, voor het koude en het warme seizoen, de scatterdiagrammen van het "verwachte verschil", respectievelijk van MC en WOL, tussen De Bilt en Eelde geplot tegen het "opgetreden verschil" tussen die twee locaties. In deze figuren zijn tevens de lineaire regressielijnen opgenomen. De correlatiecoëfficiënten behorend bij deze regressielijnen zijn $0,506$ resp. $0,545$. Beide correlaties zijn zeer significant dus de verwachte verschillen verklaren wel enigszins de opgetreden verschillen.

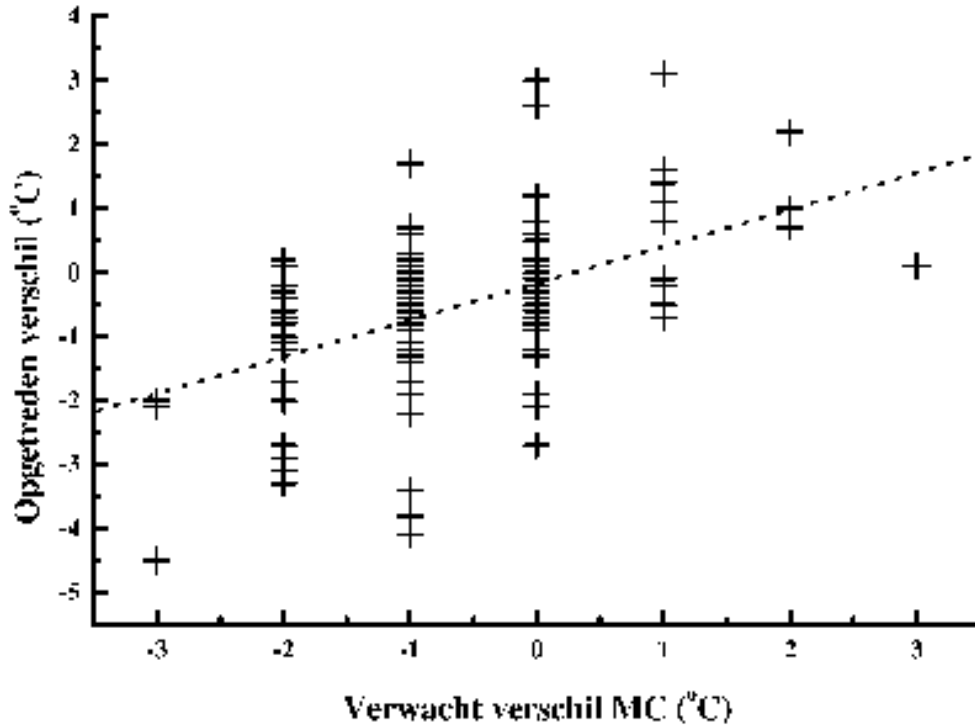


Figuur 6. Als figuur 3 maar dan voor de periode van figuur 4.

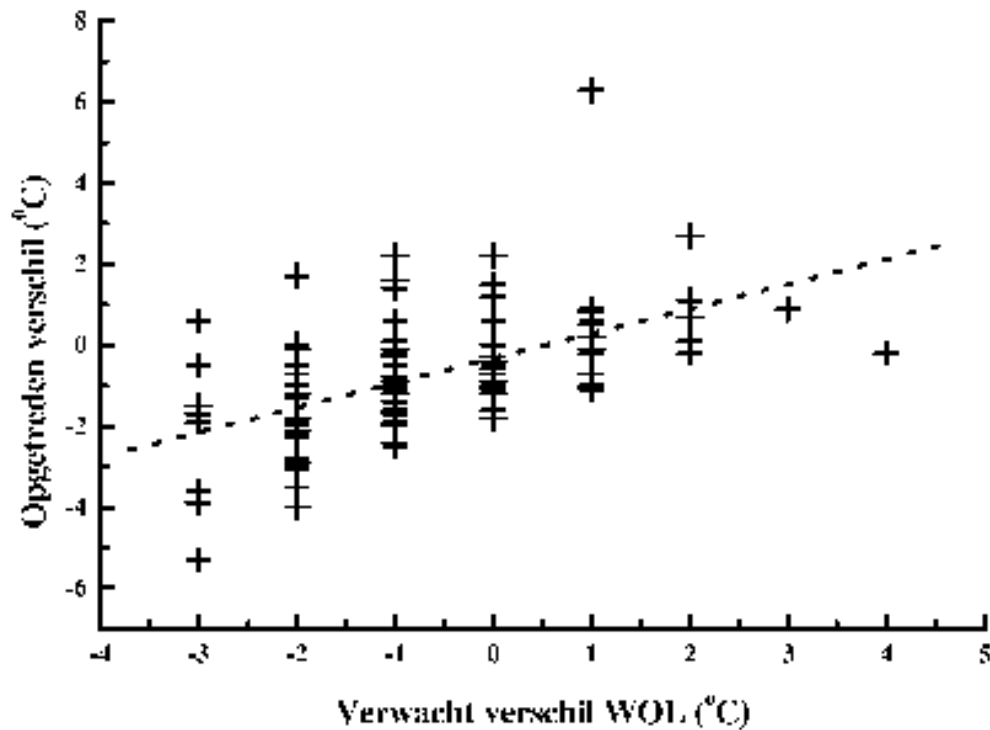
Door het "opgetreden verschil" af te trekken van het "verwachte verschil" ontstaat weer een reeks van fouten in de verwachting voor de temperatuurverschillen tussen De Bilt en Eelde. De Bias en Std van deze fouten bedroeg in het koude seizoen $-0,12$ °C en $1,15$ °C. In het warme seizoen vonden we hiervoor $0,03$ °C en $1,43$ °C. De fout in de "verwachte verschillen" is dus bijna even groot als de totale spreiding van de verschillen.

Discussie en conclusies

Zoals gezegd zijn, de hiervoor gemelde, Bias-waarden in de meeste gevallen niet hinderlijk voor het gebruik van de verwachtingen door de gebruiker. Van belang is nog wel om na te gaan of het echte Bias- en zijn of dat ze het gevolg zijn van toeval. Immers ook als we N (139 of 103) toevalsgetallen met gemiddelde nul (en standaarddeviatie s) middelen zullen we niet exact nul vinden maar een gemiddelde dat enigszins van nul afwijkt.



Figuur 7. Scatterdiagram van het opgetreden verschil in maximumtemperatuur tussen Eelde en De Bilt versus het door MeteoConsult verwachte verschil in maximumtemperatuur voor de periode van figuur 1. De gestippelde lijn is de lineaire regressielijn.



Figuur 8. Als figuur 7 maar dan voor de verwachtingen van Weer Online in de periode van figuur 4.

Om te toetsen of de afwijking zo groot is dat hij vrijwel zeker niet op toeval kan berusten gebruiken we de t-toets. Voor die t-toets delen we de gevonden Bias door de eveneens uit de steekproef berekende standaarddeviatie (Std) en vermenigvuldigen dit geheel met de wortel uit het aantal waarnemingen N. Het resultaat volgt een t-verdeling met N-1 vrijheidsgraden. Tabel 1 bevat de absolute waarde van de op deze manier berekende t-waarden van de hiervoor vermelde Bias-waarden en de bijbehorende Std.

	Koude seizoen		Warme seizoen	
	Eelde	De Bilt	Eelde	De Bilt
WNI	3.41*	5.26*	2.00	2.38
WOL	4.19*	3.21*	1.03	1.04
MC	1.23	0,06	1.34	1.28
KNMI	1.18	3.26*	2.59	2.14
Gem	3.16*	3.69*	2.04	1.91
Pers	0.27	0.12	0.59	0.60
t(99%)	2.61	2.61	2.63	2.63

Tabel 1. Resultaten voor de t-toets, van de Bias, voor alle verwachtingen en twee seizoenen. De kritieke (99%) waarde voor de t-toets is vermeld in de laatste rij. De significante t-waarden zijn gemarkeerd met een *.

In de tabel is tevens de grenswaarde vermeld die deze absolute t-waarde in 99% van alle gevallen niet overschrijdt als de berekende Bias niet afwijkt van nul oftewel de kans op t-waarde groter dan deze grens als gevolg van toeval is 1%. De Bias-waarden die vrijwel zeker niet op toeval berusten zijn in de tabel gemarkeerd met een *. Dit betreft uitsluitend Bias-waarden in het koude seizoen. WNI, WOL en KNMI moeten dus hun verwachtingsmethoden voor het koude seizoen gaan onderzoeken op Bias.

Veel belangrijker is uiteraard de vraag of de standaarddeviaties die vermeld zijn ook echt verschillend zijn. Immers ook als de onderliggende methoden resulteren in verwachtingen met vrijwel dezelfde standaarddeviatie dan kan de schatting van deze standaarddeviatie als gevolg van toeval nog variëren. Er is echter niet zoals bij de Bias een voorkeurswaarde waartegen we kunnen (willen) toetsen. De standaarddeviaties kunnen alleen onderling worden vergeleken en op basis van die verschillen moet worden geconcludeerd of die verschillen zo groot zijn dat het niet op toeval kan berusten. Als echter nu alle providers met alle andere providers gaan vergelijken en dat ook nog voor twee seizoenen en twee stations dan leidt dit tot een brij aan getallen waar mogelijk altijd wel iets bijzonders (significants) bij kan zitten. Daarom is besloten om per locatie en seizoen alleen de hoogste waarde met de laagste waarde te vergelijken en de conclusie te beperken tot de algemene conclusie "Er is wel/geen significant verschil tussen de gevonden standaarddeviaties". Vergelijking van standaarddeviaties (of varianties) wordt vrijwel altijd gebaseerd op de F-toets. Daarbij worden de twee varianties (kwadraat van de standaarddeviaties) op elkaar gedeeld met de grootste altijd in de teller en het resultaat wordt vergeleken met 99% (of 95%) waarde van de bijbehorende F-verdeling. De F-verdeling die van toepassing is wordt gekozen op grond van het aantal vrijheidsgraden n (steekproefgrootte behorend bij de teller minus 1) en m (steekproefgrootte behorend bij noemer minus 1). In dit geval zijn de steekproefgroottes van noemer en teller gelijk respectievelijk 139 en 103 in het koude respectievelijk het warme seizoen.

Als we de F-waarde echter berekenen uit de grootste en kleinste waarde uit vier mogelijkheden zijn we niet eerlijk bezig. Deze verhouding zal altijd hoger uitvallen dan verhouding tussen twee willekeurige. De 95% (en 99%) waarde van de bijbehorende F-toets

moeten dus verhoogd worden. Met behulp van het statistisch pakket R is een dergelijke situatie gesimuleerd en bij benadering vastgesteld hoeveel hoger die grenzen gekozen moeten worden. In tabel 2 zijn de resultaten samengevat. De tabel bevat zowel het 95% punt van de gewone F-verdeling alsmede de geschatte 95% en 99% punten uit de simulatie. Wederom zijn significante waarden gemarkeerd met een *.

	Locatie	Providers	F-waarde	F(0.95)	Sim(0.99)	Sim(0.95)
Koude seizoen	Eelde	KNMI/MC	1.91*	1.40	1.72	1.56
	De Bilt	KNMI/MC	1.84*	1.40	1.72	1.56
Warme seizoen	Eelde	MC/WOL	1.56	1.48	1.89	1.68
	De Bilt	KNMI/MC	1.16	1.48	1.89	1.68

Tabel 2. Resultaten voor de F-toets voor de standaarddeviaties uit figuur 3 en 6. De significante F-waarden zijn gemarkeerd met een *.

Uit deze tabel blijkt dat in het koude seizoen de verhouding tussen de varianties van hoogste en laagste provider zo hoog is dat het onwaarschijnlijk is dat dit door toeval tot stand komt. Oftewel in het koude seizoen was er een significant verschil in kwaliteit tussen de providers. In het warme seizoen is de verhouding MC/WOL wel hoog maar bij vergelijking met de gesimuleerde F-waarden zeker niet significant. Voor het warme seizoen is er dus geen aanleiding om aan te nemen dat de onderliggende kwaliteit van de verschillende providers verschillend is.

Verder valt het in de figuren 3 en 6 op dat het eenvoudigweg middelen van de verwachtingen van de vier providers schijnbaar niet leidt tot substantieel betere verwachtingen. Het gemiddelde presteert slechts marginaal beter dan de beste provider over de gegeven periode. Aangezien we echter vooraf niet weten welke provider de beste is, is het middelen van verwachtingen dus vermoedelijk wel nuttig.

Bij de regionalisatie werd een correlatie tussen verwacht verschil en opgetreden verschil tussen de locaties De Bilt en Eelde gevonden van omstreeks 0,50. De fout in de verwachte verschillen is echter bijna even groot als de spreiding in die verschillen. Het nut van de regionalisatie is dus marginaal. Als dat al geldt voor de afstand De Bilt-Eelde dan moet men twijfelen aan een regionalisatie op nog kleinere schaal. Wel dient men te beseffen dat niet alleen de fysieke afstand bepalend is. Voor locaties met sterk afwijkend lokaal klimaat kan regionalisatie nuttig zijn. Echter alvorens hiertoe over te gaan moet men eerst dit soort locaties identificeren.

Samenvattend kunnen we dus concluderen:

- In het afgelopen koude seizoen vertoonden de verwachtingen van drie van de vier providers een significante maar niet storende Bias. In de warme periode was geen Bias aantoonbaar.
- In de koude periode was er een duidelijk verschil in de standaarddeviaties van de fout in de verwachtingen en dus een aantoonbaar verschil in kwaliteit. In de warme periode is dat kwaliteitsverschil verdwenen. Deze conclusie is strikt genomen alleen geldig voor de verwachtingen die om 09:00 uur worden geëxtraheerd. Voor andere tijdstippen kan dat anders liggen.
- Regionale temperatuurverwachtingen zijn ook bij de verwachting voor de volgende dag mogelijk nog te hoog gegrepen. De standaarddeviatie van de opgetreden verschillen was echter, vergeleken bij klimatologie, aan de kleine kant. De verwachting voor de verschillen dus extra moeilijk.

Slotopmerkingen

Zoals reeds eerder is gezegd dienen vergelijkende verificaties gebaseerd te zijn op een behoorlijk lang tijdvak met gegevens. De hier gepresenteerde resultaten voldoen, mijns inziens, aan deze voorwaarde voor zover het gaat om conclusies met betrekking tot de tijdvakken die zijn geverifieerd. Maar ook hier geldt "Resultaten uit het verleden geven geen garantie voor de toekomst". Wel is het zo dat men door steeds nieuwe tijdvakken te verifiëren een beeld kan krijgen omtrent de ontwikkeling van de kwaliteit van de verwachting van de betrokken providers en op basis hiervan ook uitspraken kan doen voor de toekomst. Het project zal daarom ook nog enige tijd worden voortgezet. Zie hiervoor ook mijn webpagina [http:// home.hccnet.nl/s.kruizinga/Meteo/Verificatie](http://home.hccnet.nl/s.kruizinga/Meteo/Verificatie).

Verder zou het erg nuttig zijn om de studie voor meer parameters uit te voeren. In de toekomst zal daarom ook de minimumtemperatuur worden geverifieerd. Nog mooier zou het zijn als bijvoorbeeld ook de neerslagkans bij de studie kon worden betrokken. Echter de verwachtingen die nu worden opgeslagen bevatten slechts bij twee providers numerieke informatie over de neerslagkans.

Ten slotte dient nog vermeld te worden dat bij deze studies uitgebreid gebruik is gemaakt van het statistische pakket R (R Development Core Team, 2006). Dit pakket bevat een keur aan hulpmiddelen voor allerlei statistische analyses en is vrij verkrijgbaar via internet. Het is echter niet erg gebruikersvriendelijk.

Literatuur

R Development Core Team (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.