Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu

Opzet voor een systematische verificatie van de KNMI waarschuwingensystematiek

Pilot: Onweer 1 februari 2010 tot 1 juli 2012

F. Koek en K. Kok

De Bilt, 2016 | Wetenschappelijk rapport; WR 2016-02

Opzet voor een systematische verificatie van de KNMI waarschuwingensystematiek

Pilot: Onweer 1 februari 2010 tot 1 juli 2012

Frits Koek en Kees Kok

Inhoudsopgave

I. Introductie

a. Context

Sinds 1998 geeft het KNMI weeralarmen en (voor)waarschuwingen uit bij verwachte grootschalige extreme weercondities. Aanvankelijk gebeurde dit voor Nederland als geheel, maar sinds februari 2010 per provincie. Met diverse (markt)partijen en weerproviders is afgesproken dat er alleen gewaarschuwd wordt voor een aantal van tevoren vastgelegde meteorologische omstandigheden die op voldoend grote schaal plaatsvinden (op een zogenaamde standaard gebiedsgrootte (SGG), zie later). De criteria en ook de systematiek zijn een aantal maal aangepast. Zo is sinds 2005 een expliciete kansinschatting van overschrijding van meteorologische criteria de basis voor het verdere proces, en is in februari 2010 de geregionaliseerde waarschuwingssystematiek, met bijbehorende kleurcodering, per provincie ingevoerd. Deze verandering was ingegeven om beter aan te sluiten bij de beleving van de gebruikers en te zorgen voor vermindering van het aantal false alarms. ("Een vernieuwd weeralarm", KNMI, 2010). De uitgifte van weeralarmen is sindsdien voor-behouden aan een speciale commissie waarin het KNMI zitting heeft naast verschillende maatschappelijke instanties.

Uitgegeven en mogelijk onterecht gemiste weeralarmen worden sinds de invoering standaard geëvalueerd, waarbij de gemiste waarschuwingen iets minder gedetailleerd worden bekeken. Deze evaluatie betreft meteorologische en niet-meteorologische elementen. Naast de beoordeling of de uitgifte meteorologisch gezien terecht was, wordt er gekeken naar de beoordeling door de 'samenleving' (media), of er sprake is geweest van ontwrichting als gevolg van het weer, of de communicatie goed is verlopen, of de infrastructuur goed gefunctioneerd heeft, e.d..

Meteorologische *verificatie* van de 'weeralarmen' is hierbij beperkt tot de vergelijking waarbij 'ja/nee gealarmeerd' voor heel Nederland vergeleken is met 'ja/nee waargenomen' ergens in Nederland (Mureau, 2005; "Een vernieuwd weeralarm" (KNMI, 2010); www.knmi.nl). Op een dergelijke categorische verificatie is veel af te dingen (Kok, 2005). Het optreden van extreme condities is vaak onzeker en is relatief zeldzaam. De mogelijkheid om de rol van de meteoroloog bij het opstellen van een verwachting voor extreme omstandigheden op waarde te schatten is daardoor eveneens beperkt. Sinds 2005 is gedurende enkele jaren in het Team Integrale Kwaliteit (TIK-team) bij de evaluaties van (bijna) weeralarm situaties de onzekerheid in de waarnemingen wèl expliciet meegenomen in de vorm van een kansschatting in plaats van 'ja/nee waargenomen'. Hierbij werd nog niet gekeken naar de precieze locatie van het betreffende fenomeen.

Om te komen tot een objectieve verificatie van de gevaarlijk weer verwachtingen moet een aantal keuzes gemaakt worden. Zo zijn er veel onzekerheden bij de beoordeling of een criterium gehaald is of niet. Deze onzekerheden hebben bijvoorbeeld te maken met het overschrijden van de betreffende meteorologische drempel (i.e. zijn er voldoende waarnemingen en wat is hun nauwkeurigheid en representativiteit?), met de grootte en

ligging van het getroffen gebied (overlapt het gebied met vooraf gedefinieerde regio's?), en de coherentie van het waargenomen fenomeen.

Aan de hand van een pilot waarin onweerwaarschuwingen in de categorie Oranje (extreem weer) en Rood (weeralarm) worden geanalyseerd wordt in dit rapport een beschrijving gegeven van een mogelijke systematische verificatie van de geregionaliseerde verwachtingen voor gevaarlijk of extreem weer. Er wordt uitgegaan van het systeem dat operationeel was van februari 2010 t/m september 2015[1], waarin waarschuwingen per provincie en in hele uurvakken werden gegeven en gecommuniceerd m.b.v. een kleurcodering (variërend van Groen (geen waarschuwing), Geel (waarschuwing voor gevaarlijk weer), Oranje (waarschuwing voor extreem weer) tot Rood (weeralarm)). Het betreft verificatie van de uiteindelijk aan het publiek, via Internet, gepresenteerde waarschuwingen. Hoewel de meteorologie slechts een deelrol speelt worden in deze pilot de waarschuwingen vergeleken met meteorologische waarnemingen. De impact die 'de maatschappij' van het extreme weer ondervindt – een leidende factor in de waarschuwingssystematiek – wordt niet in de verificatie meegenomen.

De keuze voor een afgebakende analyse (onweerwaarschuwingen met code Oranje of Rood) is ingegeven door de aanwezigheid van goed dekkende waarnemingen (i.t.t. bij sommige andere predictands) in de vorm van tellingen van alle ontladingen. Desondanks moest er een aantal aannames gemaakt worden. Deze worden samen met de verificatieresultaten beschreven, en het rapport geeft tot slot enige discussiepunten en aanbevelingen voor mogelijke verbeteringen.

b. Het waarschuwingssysteem van februari 2010 t/m september 2015

Het proces om te komen tot een extreem weer waarschuwing of weeralarm, bevat een aantal stappen (Fig. I.1). De basis is de inschatting van de kans op het overschrijden van de gedefinieerde meteorologische criteria via een consensus-protocol door de meteorologen op dienst (eventueel in samenspraak met geraadpleegde experts). Wordt deze overschrijdingskans op minstens 60% geschat dan volgt een consultatie met de weerproviders, en wordt al of niet besloten tot code Oranje. Wordt na deze consultatie de (consensus) kans op tenminste 90% geschat, dan wordt het complete weeralarmteam bij elkaar geroepen, met daarin een aantal maatschappelijke partijen die een (subjectieve) inschatting van de maatschappelijke impact maakt alvorens eventueel over te gaan op het afgeven van een code Rood. Voor code Geel geldt een heel ander traject met andere (in sommige gevallen meteorologisch minder goed gedefinieerde) criteria. Dit blijft in deze analyse buiten beschouwing.

Code Oranje wordt gegeven bij twee situaties: 1. bij een kans tussen 60% en 90% en 2. bij een kans ≥ 90% maar lage geschatte impact. Code Rood wordt gegeven bij een kans ≥ 90% waarbij de impact hoog wordt ingeschat. Situaties met een hoge impact maar een geschatte kans <90% krijgen dus maximaal code Oranje. Voor code Oranje en Rood zijn de meteorologische criteria identiek, en speelt, naast impact, dus alleen de ingeschatte kans op overschrijding van deze criteria een rol.

[1] Hierna wordt een systematiek gehanteerd zoals beschreven in "Herijking Waarschuwingssystematiek 2015".

De weerfenomenen waarvoor criteria zijn opgesteld zijn (zie Fig. I.2) Windstoten, Onweer, Regen, Gladheid/sneeuwval, Zicht, Extreme hitte, Wind- en waterhozen. Alleen voor Windstoten, Onweer, Regen en Gladheid/sneeuwval zijn alle codes (Groen, Geel, Oranje en Rood) gedefinieerd, voor de overige elementen alleen code Groen of Geel.

De kansinschattingen hebben ook een relatie met de tijd voorafgaand aan de gebeurtenis waarop een waarschuwing wordt afgegeven. De zekerheid omtrent een extreme gebeurtenis neemt toe als de tijd tot de gebeurtenis relatief kort is. Daarom wordt een code Rood in de regel later afgegeven dan een code Oranje en nooit eerder dan 12 uur voor het tijdstip van het verwachte fenomeen (zie Fig. I.2).

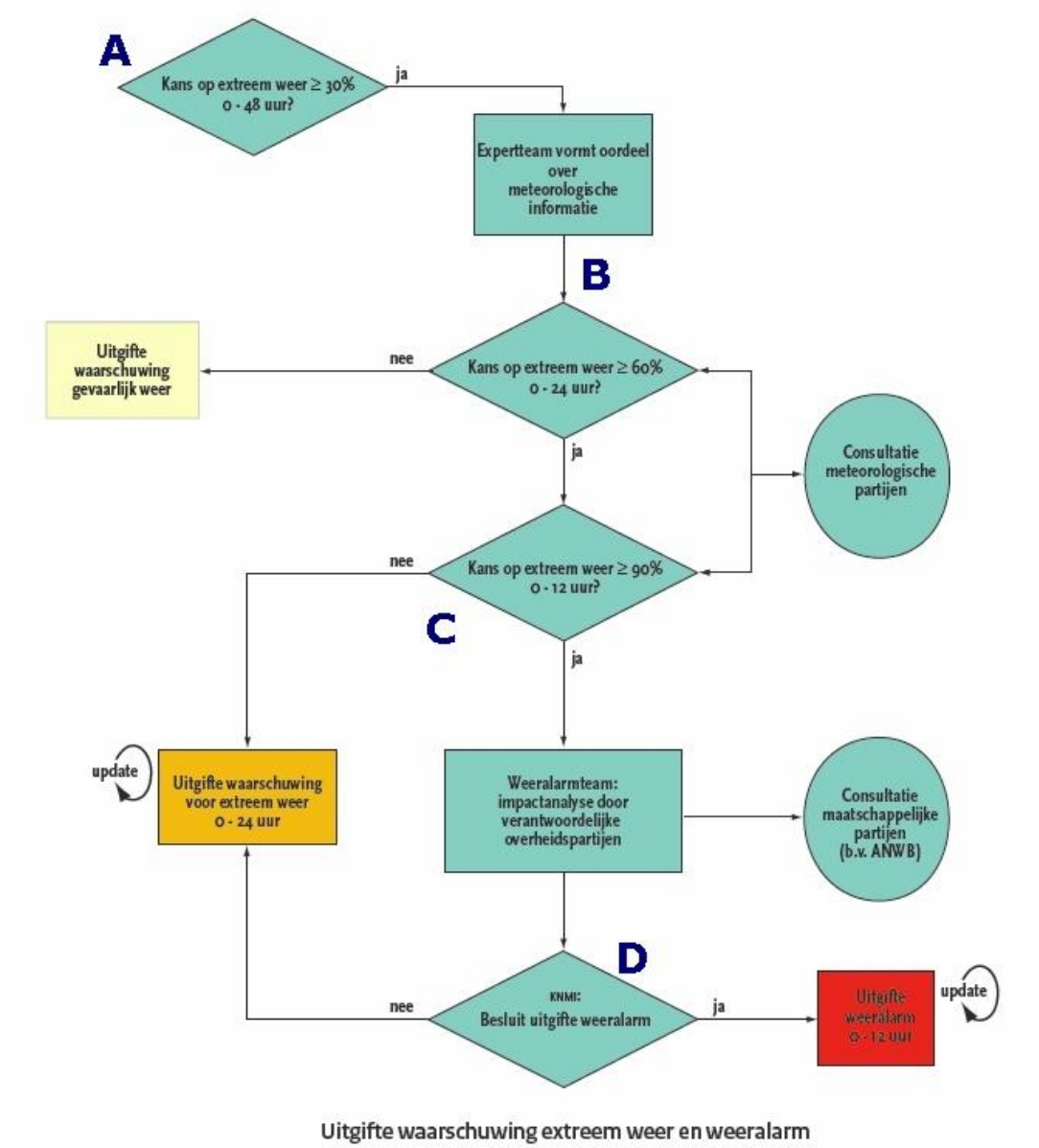
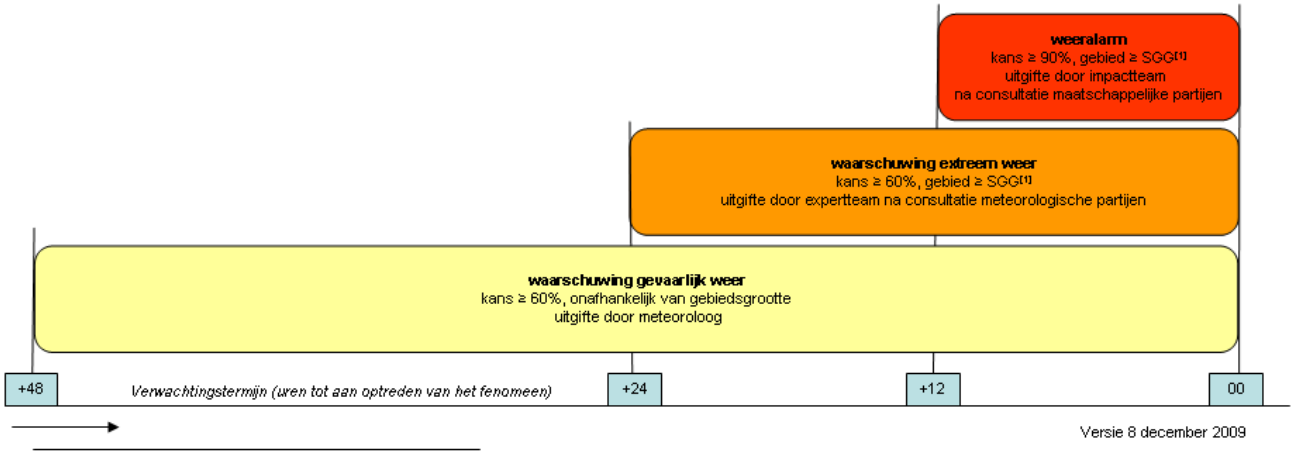


Fig. I.1. Stroomschema uitgifte waarschuwingen voor extreem weer en weeralarmen. Bron: 'Uitgifteproces Waarschuwingen en weeralarmen' (12 januari 2010). Geldig t/m september 2015.

De waarschuwingen voor extreem weer en weeralarmen worden afgegeven wanneer het fenomeen is opgetreden of wordt verwacht in een gebied tenminste ter grootte van de zogenaamde Standaard Gebieds Grootte (SGG). Deze is gedefinieerd als een *gebied van minimaal 50km x 50km of een coherente band met een lengte van tenminste 50km boven het Nederlandse vasteland*. Voor de verificatie van onweer is een coherente band van ten minste 50km lengte equivalent met een gebied van 50 bij 50 km.

De meteorologische indicatoren en de bijbehorende overschrijdingscriteria waarvoor een weerwaarschuwing wordt afgegeven staan in Fig. I.2. Voor onweer geldt dat er 500 ontladingen per 5 minuten binnen een SGG moeten optreden, voordat een code oranje of rood wordt afgegeven. Een voorbeeld van de presentatie bij een uitgegeven waarschuwing voor (in dit geval) zware neerslag staat gegeven in Fig. I.3.

Fenomeen	Criterium waarschuwing gevaarlijk weer	Criterium waarschuwing extreem weer & weeralarm
Gladheid en sneeuwval	▪ hinder door hagel, op- of aanvriezing of bevriezing van natte weggedeelten ▪ accumulatie sneeuw 0-3 cm/uur	▪ accumulatie sneeuw >3cm/uur of >10 cm/6 uur ▪ sneeuwval of driftsneeuw met wind >40 km/uur ▪ gladheid door ijzel of ijsregen
Onweer	▪ >1 ontlading in 5 minuten, al dan niet met hagel	▪ >500 ontlading in 5 minuten, al dan niet met hagel
Regen	▪ hinder voor het wegverkeer (uitsluitend na melding VIF-partners) ▪ >30 mm/uur	▪ >75 mm in 24 uur
Wind- en waterhozen	▪ waarneming wind- of waterhoos	▪ geen
Windstoten	▪ >75 km per uur	▪ >100 km per uur ▪ >120 km per uur (winterperiode, kuststrook)
Zicht	▪ ≤ 200m ▪ ≤ 50m ▪ ≤ 10m	▪ geen



[1] SGG: Standaard Gebieds Grootte.
Fenomeen opgetreden of verwacht in een gebied van minimaal 50x50 km of een coherente band met een lengte van tenminste 50 km boven het Nederlandse vasteland.

Fig. I.2. Meteorologische criteria die worden getoetst voor de afgifte van een gevaarlijk weer waarschuwing, een waarschuwing voor extreem weer of een weeralarm. Code Geel mag maximaal 48 uur voor het verwachte fenomeen uitgegeven worden, voor code Oranje en Rood is dit resp. 24 en 12 uur. Bron: Folder 'KNMI waarschuwingen Nederland. De veranderingen vanaf 01-02-2010 (mei 2010).' Vanaf oktober 2015 zijn andere criteria geldig.

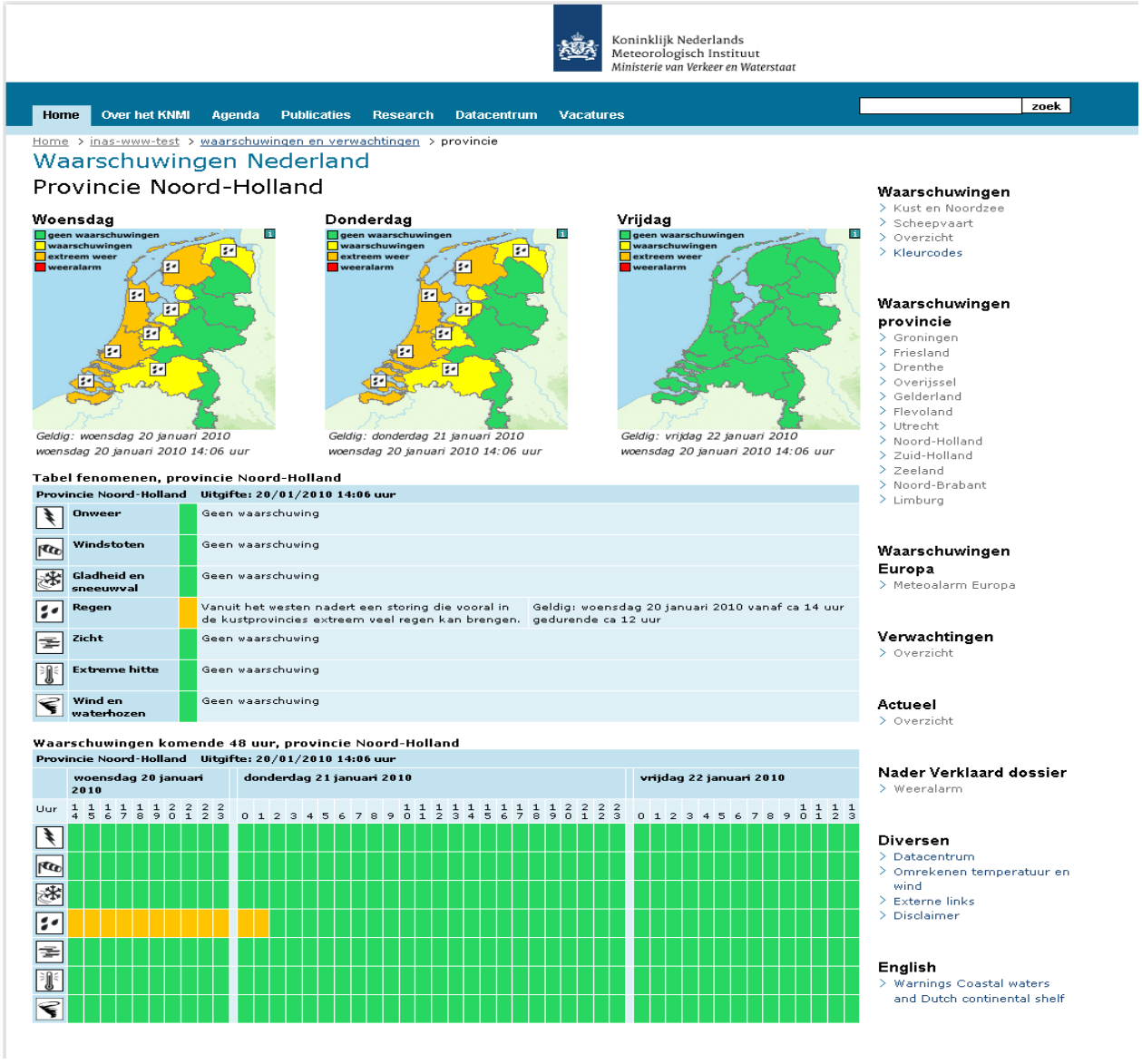


Fig. I.3. Presentatie van de waarschuwingen uitgegeven op 20 januari 2010 om 14 uur. Een overzicht van alle codes voor ‘vandaag’, ‘morgen’ en ‘overmorgen’ voor alle provincies staat in de bovenste 3 panels. In het onderste panel staan de uurlijkse kleurcodes voor de eerste 48 uur vanaf 14 uur voor Noord Holland.

c. Doel en begrenzing van dit rapport

Doel van iedere verificatie is, naast het weergeven en kwantificeren van resultaten uit het verleden, het identificeren van eventuele tekortkomingen en het aangeven van mogelijke verbeteringen. Voor het weerwaarschuwingssysteem (Fig. I.1) dient verificatie (of evaluatie) van ieder van de genoemde deelprocessen A t/m D plaats te vinden. Vanwege de beperkte beschikbaarheid van gegevens is verificatie alleen mogelijk voor de kansschattingen van de meteoroloog (punt A in Fig. I.1) en de op Internet gepubliceerde kleurcodes die echter mede afhankelijk zijn van de impact inschatting door het Weeralarmteam (punt D). Deze analyse beperkt zich tot de verificatie van D tegen het opgetreden weer (en niet tegen de impact). Kok et al (2011b) hebben een initiële verificatie van A beschreven. Indien de totstandkoming van het consensus-proces (stap B en C) afdoende wordt gedocumenteerd kunnen ook deze onderdelen worden geverifieerd.

Het is aan te bevelen om een verificatie van alle meteorologische grootheden in de gevaarlijk weer systematiek uit te voeren. Vanwege de beschikbaarheid van goed dekkende, objectieve, kwantitatieve data wordt in dit rapport aandacht besteed aan *Onweer*. Deze verificatie kan de ontwikkeling van andere verificatie-methodieken helpen.

We beperken ons in dit rapport tot de kleurcodes Rood en Oranje omdat dit de extreem weer waarschuwingen betreft. Code Geel (meer dan 1 ontlading) is weliswaar goed gedefinieerd maar is hier genegeerd omdat er nogal eens valse detecties van ontladingen voorkomen (zgn. bliksemontladingen bij heldere hemel).

II. Onzekerheden bij de waarschuwingssystematiek

In dit hoofdstuk wordt gewezen op een aantal onzekerheden en ambivalenties in het gehanteerde waarschuwingssyteem. Deze onzekerheden maken een vertaling van verwachte weersextremen naar een kleurcode per provincie per uur niet in alle gevallen geheel duidelijk. Dit heeft consequenties voor de manier waarop geverifieerd moet worden en voor de interpretatie van de resultaten.

a. De omschrijving van de standaard gebiedsgrootte (SGG)

Een van de onduidelijkheden betreft de ligging van de Standaard Gebieds Grootte, onder andere of het SGG gebied geheel binnen de landsgrenzen ("boven het Nederlandse vasteland") moet liggen. Wat te doen bij de Noordzee, Waddenzee of IJsselmeer, etc.? Strikte toepassing van de definitie zal de kans op een weeralarm in bijna de helft van Nederland sterk verlagen.
Ook de oriëntatie van het (vierkante) SGG gebied kan leiden tot onduidelijkheden. Vrijwel altijd is het SGG 'oost-west georiënteerd' (i.e. met 2 zijden langs breedtecirkels). Dit betekent dat de signalering en de intensiteit van het fenomeen in principe iets kunnen worden onderschat (zie hfdst. III).
Voor de vertaling van SGG naar provincies zie hoofdstuk III.

b. Verdeling van kansen binnen een tijdsperiode

Als een meteoroloog een kansschatting van bijvoorbeeld 60% maakt voor het optreden van een bepaald fenomeen (*event*) en er code oranje afgegeven wordt, dan wordt vaak impliciet bedoeld dat *binnen een bepaald tijdsinterval* de kans van optreden 60% is. De waarschuwingscode voor ieder uur in dat interval wekt de suggestie dat de kans van optreden op ieder uur (tenminste) 60% is. Dit is een voor de hand liggende interpretatie voor de gebruiker, die immers geen andere informatie heeft. Deze interpretatie zou in de verificatie per uur dus een overforecasting te zien moeten geven.

c. Gecombineerde kansen van meerdere predictands

Bij de interpretatie van de verificatieresultaten moet tenslotte nog rekening gehouden worden met de combinatie van events, zoals bijv. onweer in combinatie met windstoten. Het is voor een meteoroloog niet altijd duidelijk wat hij / zij geacht wordt te doen als de kans op overschrijding voor elk van beide fenomenen kleiner is dan het kanscriterium, maar de gecombineerde kans groter. In sommige gevallen wordt dan toch een waarschuwing afgegeven voor één of voor beide elementen. Criteria voor gecombineerde fenomenen zijn niet opgenomen in de waarschuwingssystematiek (Fig. I.2). Afgeven van een waarschuwing is wellicht conform de beleving van het publiek, maar leidt kwantitatief gezien tot een overschatting van het fenomeen onweer.

III. Onzekerheden bij de gebruikte waarnemingen

Met behulp van waarnemingen van ontladingen wordt per provincie (of meerdere tegelijk) bepaald of aan het criterium voldaan is of niet. Omdat de SGG in het algemeen niet samenvalt met provinciegrenzen moet, als het vereiste aantal ontladingen gehaald is, nog een kleurtoewijzing plaats vinden naar de onderliggende provincies. Hiervoor worden twee methoden gehanteerd (zie ook hoofdstuk IV) die zullen worden vergeleken.

a. Gebruikte waarnemingen

De onweerwaarnemingen zijn afkomstig van het detectiesysteem FLITS (Flash Localisation by Interferometry and Time of Arrival System), dat een upgrade is van het SAFIR (Surveillance et Alerte Foudre par Interférométrie Radioélectrique) systeem (Wessels 1998, Noteboom 2006). In dit rapport wordt een module (van Westrhenen, pers comm.) toegepast waarin op een historische dataset van FLITS het maximaal aantal ontladingen in 'oost-west georiënteerde' gebieden van 50 x 50 km berekend wordt in disjuncte stappen van 5 minuten. Per uur zijn er dus 12 waarnemingen. De ruimtelijke resolutie van de module is 0.001 graad (ca 70 à 100m).

In overeenstemming met het onweercriterium wordt geen onderscheid gemaakt tussen horizontale en verticale ontladingen.

b. Implicaties voor de verificatie

- Het gebruik van vierkanten die 'oost-west georiënteerd' zijn veroorzaakt een (kleine) **onderschatting** van het aantal opgetreden gebeurtenissen. Als de oriëntatie van het onweergebied in Fig. III.1 O-W geweest zou zijn in plaats van NW-ZO zouden er in het SGG minder ontladingen zijn geteld. Situaties met een O-W of N-Z patroon van de ontladingen hebben de grootste kans om ten onrechte niet geteld te worden als extreem weer of weer-alarmsituatie.

Fig III.1 Voorbeeld van een situatie waarin de ontladingen in een 5-minuten interval geconcentreerd zijn langs een diagonaal van het 'oost-west georiënteerde' 50x50 km vierkant.

De meteoroloog baseert de beoordeling van de actuele onweerssituatie op het aantal ontladingen in een vierkant van 50x50km dat in ‘oost-west richting’ georiënteerd is. Hierdoor ontstaat een verschillende beoordeling van de zwaarte van het onweer afhankelijk van de oriëntatie van het weerfenomeen. Dit kan opgelost worden door een willekeurige oriëntatie van het gebied toe te staan. Dit betekent dat vierkanten van 50x50 km gezocht worden waarin aan het criterium voldaan is in een gebied dat een factor $\pi/2$ groter is dan het SGG. Ook een cirkelvormig gedefinieerd SGG kan dit verhelpen.

- De mismatch tussen gedefinieerde SGGs en provinciegrenzen kan bijv. worden verholpen door gebruik van ‘affected areas’: zodra in een gebied ter grootte van het SGG aan het criterium voldaan is, wordt voor alle provincies die een overlap met het SGG hebben verondersteld dat het criterium is overschreden (zie hoofdstuk IV). Als het onweer zich beperkt tot een zeer klein gebiedje dan kan het gebied opgespannen door gebieden ter grootte van het SGG waarin het criterium gehaald is veel groter zijn dan 50x50km. In het voorbeeld van Fig. III.2 strekt de ‘affected area’ met meer dan 100 ontladingen per 5 minuten boven Noord Brabant zich uit over de provincies Zeeland, Zuid Holland, Utrecht en Gelderland. In de verificatie betekent dat bij het hanteren van ‘affected areas’ het aantal waargenomen events per provincie toeneemt en het aantal false alarms afneemt.

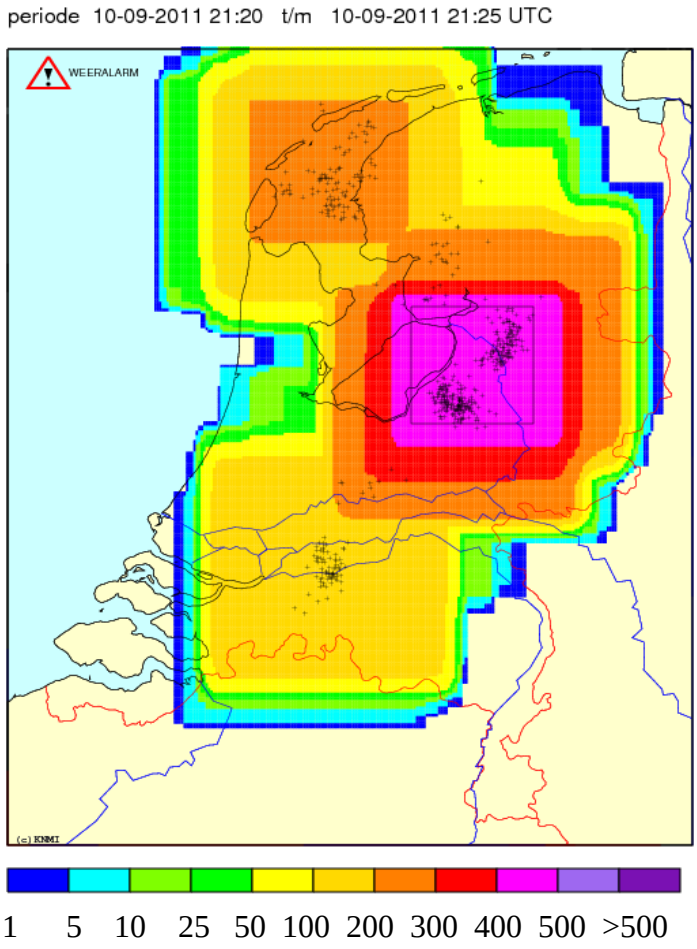


Fig. III.2. Voor een serie drempelwaarden van aantallen ontladingen (aangegeven onder de figuur) is voor ieder punt in NL gekeken of dat punt binnen een 50x50km vierkant ligt waarbinnen dat aantal overschreden is. Het betreft een 5-minuten interval op 10 september 2011.

- De bliksemtellingen zijn beschikbaar in disjuncte stappen van 5 minuten in plaats van in ‘lopende’ 5 minuten intervallen. Dit veroorzaakt ook een (kleine) **onderschatting** van het aantal opgetreden gebeurtenissen.

IV. Opzet van de verificatie-methodiek

Om het effect van de in het vorige hoofdstuk beschreven onzekerheden te verkennen wordt in deze verificatie een aantal aannames gemaakt. Dit betreft onder andere de standaard gebiedsgrootte (SGG) (definitie bij landsgrenzen, provinciegrenzen of bij land-water overgangen; definitie van de coherente band).
Een aantal opties wordt beschreven en doorgerekend. Dit dient als basis voor verdere discussie en aanscherping.

De uitgevoerde verificatie is in essentie een categorische vergelijking van het afgeven van een code Oranje/Rood (ja/nee) versus het halen van het meteorologische criterium (ja/nee). De meest directe en zuivere manier is om voor alle provincies en voor alle (forecast) uren te kijken of een waarschuwing terecht was *(hit*) of niet (*false alarm*) en of er terecht of niet niet is gewaarschuwd (resp. *correct rejection* en *miss*). Bij de beoordeling hiervan kunnen waarnemingen op verschillende manieren worden geinterpreteerd. Een waargenomen event (dwz waarneming = ja) kan worden vastgesteld als het centrum van de onweersactiviteit *in* de betreffende provincie ligt, maar ook als de onweersactiviteit een provincie '*raakt*', i.e. als een deel van de provincie onderdeel uitmaakt van het gebied van 50km x 50km waarin het criterium overschreden is. De eerste definitie is veel strenger dan de tweede. Om een indicatie te krijgen van de consequentie van deze aanname wordt de verificatie met beide opties uitgevoerd: respectievelijk *'Midden van SGG'* en *'Corners van SGG'* genoemd omdat we hierbij gekeken hebben of het centrum respectievelijk de hoekpunten van het SGG-gebied binnen een bepaalde provincie liggen. Bij de optie '*Corners van SGG'* worden daarom SGG-gebieden meegenomen waarbij het centrum binnen een 25km zone rond Nederland ligt (zie Fig. IV.1).

Fig. IV.1 De provincies en de 25km zone rond Nederland.

Bovenstaande is uitgevoerd zowel afhankelijk als onafhankelijk van de forecasttijd. Daarnaast is ook een verificatie uitgevoerd waarbij het locatie-criterium is losgelaten en de waarschuwingen (per provincie per uur) geverifieerd zijn tegen 'ergens in Nederland' opgetreden of niet. Deze laatste optie maakt een vergelijking mogelijk met de vroegere weeralarmsystematiek van vóór de regionalisatie. Voor een goede vergelijking wordt hier ook een zone van 25km buiten de Nederlandse grens en kustzone bij betrokken.

Bij alle bovenstaande opties worden tenslotte ook gevoeligheidsanalyses uitgevoerd met een lagere drempelwaarde voor het waargenomen aantal ontladingen (300 ipv 500 ontladingen per 5 minuten), om een indruk te krijgen van het gewicht van bijna-hits in de resultaten. De resultaten hiervan staan in Appendix A1.

Een samenvatting van de manieren waarop geverifieerd is staat in Tabel IV.1.

Verificatie	Midden van SGG	Corners van SGG
tegen 500 / 300 ontl.'n per 5 min: per provincie, per uurvak	VI.1.A / App.A1.A	VI.1.B / App.A1.B
tegen 500 ontl.'n per 5 min: per provincie, als functie van fc tijd	VI.2	App. A2
tegen 500 / 300 ontl.'n per 5 min: ergens in NL, per uurvak	VI.1.A / App.A1.A	VI.1.B / App.A1.B
tegen 500 ontl.'n per 5 min: ergens in NL, als functie van fc tijd	VI.2	App. A2

Tabel IV.1. Overzicht van de uitgevoerde verificaties met de hoofdstukken en appendices waarin die beschreven zijn.

V. Metriek

a. Gebruikersscores

De kwaliteit van de waarschuwingen wordt op een aantal manieren gepresenteerd. De 'kale' representatie van de 'waarschuwing,waarneming' paren wordt gegeven in de vorm van 2x2 contingentie tabellen (Tabel V.1) voor alle provincies apart, voor Nederland als geheel (aangegeven met NL) en voor de vereniging van alle provincies (dwz alle provincies gezamenlijk, ALL). Daarnaast wordt voor deze 3 indelingen (alle afzonderlijke provincies, NL en ALL) zowel grafisch als in getalvorm een aantal scores getoond: de Probability of Detection (POD), de False Alarm Ratio (FAR) en - daaruit volgend - de Bias (B, ook wel Bias ratio genoemd) en de Critical Success Index (CSI, ook wel Threat Score genoemd). Deze scores kunnen gezien worden als 'gebruikersscores', omdat ze direct te relateren zijn aan de te verwachten schade of kosten als gevolg van gemiste events of onterecht gegeven waarschuwingen. Hoe waardevol of nuttig de waarschuwingen zijn geweest voor gebruikers hangt vervolgens af van de individuele omstandigheden waarvoor ze gebruikt worden, bijv. van de gevoeligheid voor misses en false alarms van de betreffende gebruiker, of van diens zgn. cost-loss verhouding, i.e. de verhouding van de kosten (costs) ter voorkoming of vermindering van schade tot de verliezen (losses) die optreden als er geen voorzorgs-maatregelen getroffen zijn (e.g. Bilham, 1922; Bijvoet en Bleeker, 1951; Katz and Murphy, 1997; Richardson, 2000).
Omdat deze 4 scores nauw met elkaar verbonden zijn kunnen ze in één plot zichtbaar gemaakt worden. Bij de presentatie van de resultaten maken we gebruik van 'POD versus FAR' plots, ook wel performance diagrams geheten (Roebber, 2009), waarin tevens isolijnen van CSI en B gegeven zijn.
De False Alarm Rate (F) wordt ook gegeven omdat deze onderdeel is van de skill scores die gepresenteerd worden in de verificatie (zie sectie V.b).

		waargenomen	
		YES	**NO**
gewaarschuwd	**YES**	A	B
	NO	C	D

Tabel V.1. 2x2 contingentie tabel. Hierbij is A het aantal hits, B het aantal false alarms, C het aantal misses en D het aantal correct rejections. Het aantal uitgegeven waarschuwingen is A+B en het aantal opgetreden gevallen boven de aangegeven drempelwaarde is A+C.

De definities van de hier gebruikte scores zijn als volgt:
Probability Of Detection geeft de fractie van alle waargenomen gevallen waarvoor terecht gewaarschuwd is:

$$POD = \frac{A}{A+C}$$

False Alarm Ratio geeft de fractie van alle waarschuwingen die onterecht waren:

$$FAR = \frac{B}{A+B}$$

False Alarm Rate geeft de fractie van alle niet waargenomen gevallen waarvoor wél (dus onterecht) gewaarschuwd is:

$$F = \frac{B}{B + D}$$

Critical Succes Index geeft de fractie van alle 'niet-triviale' gevallen waarvoor terecht gewaarschuwd is:

$$CSI = \frac{A}{A + B + C}$$

Bias geeft de verhouding van gewaarschuwde en waargenomen aantal gevallen:

$$BIAS = \frac{A + B}{A + C}$$

Een bias van groter dan 1 duidt op overforecasting, van kleiner dan 1 op underforecasting.

Voor een perfect verwachtingssysteem (i.e. B = C = 0) zijn de FAR en de F gelijk aan 0, en de POD, CSI en B gelijk aan 1. Het verwachte gedrag van deze scores is sterk afhankelijk van het opgetreden aantal gevallen, of algemener, van de 'klimatologie' van het fenomeen (Halsey, 1995; Kok, 2000). Bij weinig voorkomende fenomenen hebben zowel de POD als de CSI de neiging om naar nul te gaan en de FAR naar 1. Dit gedrag is voor alle niet-perfecte verwachtingssystemen zichtbaar. Daarom is het moeilijk om met behulp van deze scores de kwaliteit van verwachtingen te vergelijken voor fenomenen die ongelijke klimatologie hebben, i.e. voor predictands die verschillend zijn, of voor dezelfde predictands maar geverifieerd op verschillende plaatsen of voor verschillende drempelwaarden of periodes waarop geverifieerd wordt.

b. skill scores

Voor de skill, de kwaliteit van de verwachtingen t.o.v. een of andere referentieverwachting, zijn de bovengenoemde 'gebruikerscores' niet informatief. Skill scores zijn 0 bij geen skill t.o.v. de referentieverwachtingen en zijn 1 bij perfecte (categorische) verwachtingen, en negatief als ze slechter zijn dan de referentieverwachtingen. Veel gebruikte referentieverwachtingen zijn random verwachtingen (met verschillende aannames over het percentage verwachte events), persistentie (de verwachting is gelijk aan de huidige waarneming) of klimatologie (meestal een langjarig gemiddelde van de predictand in het zelfde jaargetijde als waarvoor de verwachtingen gemaakt zijn). De keuze van een skill score voor *rare events* is niet triviaal. Een vaak gebruikte score is de Hanssen Kuipers Score (HKS, Hanssen en Kuipers, 1965) die de gunstige eigenschap heeft dat de verwachtingswaarde nul is voor alle 'unskillful' verwachtingen (*equitability*; Gandin and Murphy, 1992), zoals verwachtingen waarbij random een van de 2 klasses gekozen wordt, of altijd dezelfde klasse. De HKS is echter minder geschikt voor de verificatie van verwachtingen van zeldzame events. Dan convergeert de score namelijk naar de POD die gemakkelijk is te optimaliseren door bijvoorbeeld onrealistisch vaak het fenomeen te voorspellen.
Daarnaast gaan de meest gebruikte traditionele scores in de limiet voor toenemende zeldzaamheid naar een zgn. niet-informatieve waarde (meestal 0), onafhankelijk van de kwaliteit van het verwachtingssysteem. Dit (ongewenste) gedrag is bijvoorbeeld te zien voor de HKS in Fig. V.1 waarin voor gebiedsneerslagverwachtingen voor waterschap Delfland het

gedrag van een aantal scores staat als functie van overschrijdingsdrempel (top panel; neerslag in tienden mm's) en observed frequency (bottom panel).
Recent zijn verschillende nieuwe categorische verificatiescores ontwikkeld die dit nadeel - de grote afhankelijkheid van de drempelwaarde van de betreffende predictand - niet hebben en die speciaal bedoeld zijn voor de bepaling van de skill van systemen in het verwachten van *rare events*. De eerste was de Extreme Depency Score (EDS; Stephenson et al., 2008; Ghelli and Primo, 2009) gevolgd door een verbeterde versie, de Symmetric Extreme Dependency Score (SEDS; Hogan et al., 2009), Omdat deze scores *base rate dependent* zijn ontwikkelden Ferro and Stephenson (2011) de Extremal Dependence Index (EDI) en de Symmetric Extremal Dependence Index (SEDI). Het niet afhankelijk zijn van de base rate (sample klimatologie) is een belangrijke eigenschap, speciaal voor het monitoren van de skill van een verwachtingssysteem over meerdere jaren. Bij het gebruiken van base rate dependent scores is men niet zeker of een verandering in de waarde van de score het gevolg is van een verandering van de skill of van de base rate.
Een nadeel van al deze nieuwe scores is dat voor een goede interpretatie van de scores de verwachtingen eerst gekalibreerd moeten worden (i.e. bias = 1). Bij de onderhavige verwachtingen zoals getoond in het volgende hoofdstuk is dit geenszins het geval. Het effect hiervan is op voorhand niet duidelijk en is nog onderwerp van onderzoek. Hetzelfde geldt voor de methodes om de verwachtingen te kalibreren.
Het gedrag van deze scores bij toenemende zeldzaamheid is te zien in Fig. V.1. Omdat het gekalibreerde verwachtingen betreft, i.e. bias (vrijwel) gelijk aan 1, zijn de EDS en SEDS (min of meer) gelijk. Voor een uitgebreide beschrijving van deze nieuwe scores en een vergelijking van hun voor- en nadelen wordt verwezen naar Ferro and Stephenson (2011).

Omdat nog veel onderzoek naar de theoretische en praktische eigenschappen van de scores nodig is (ECMWF-TAC, 2010) maken we in deze studie vooralsnog geen keuze en presenteren we naast de HKS drie van de nieuwe scores: de SEDS, EDI en SEDI. Kalibratie van de verwachtingen was met de beschikbare data niet mogelijk. Daarom presenteren we de skill scores zonder commentaar en concentreren we ons in de bespreking van de resultaten op de gebruikersscores, m.n. de POD en FAR.

De definities van de gebruikte (skill) scores zijn als volgt:

Hanssen-Kuipers Score:

$$HKS = \frac{AD - BC}{(A + C)(B + D)} = POD - F$$

Symmetric Extreme Dependency Score:

$$SEDS = \frac{\ln\left(\frac{A+B}{N}\right) + \ln\left(\frac{A+C}{N}\right)}{\ln\left(\frac{A}{N}\right)} - 1$$

hierbij is N het totaal aantal gevallen, A+B+C+D.

Extremal Dependence Index:

$$EDI = \frac{\ln(F) - \ln(POD)}{\ln(F) + \ln(POD)}$$

Symmetric Extremal Dependence Index:

$$SEDI = \frac{\ln(F) - \ln(POD) - \ln(1-F) + \ln(1-POD)}{\ln(F) + \ln(POD) + \ln(1-F) + \ln(1-POD)}$$

c. Het uitsluiten van de meest triviale gevallen

Bij skill scores wordt een succesvolle detectie van een situatie waarin een fenomeen *niet* optreedt (cel D in de contingentietabel) ook meegenomen bij de beoordeling van de kwaliteit van een verwachtingssysteem, i.t.t. bij de gebruikersscores waarin alleen cellen A, B en C gebruikt worden. Bij rare event forecasting bevat deze cel echter normaliter een groot aantal 'triviale' gevallen. Om te voorkomen dat de scores gedomineerd worden door deze gevallen, en daarmee relatief ongevoelig worden voor veranderingen in de kwaliteit in niet-triviale omstandigheden, i.e. meestal de belangrijkere gevallen, is stratificatie van de dataset noodzakelijk (Murphy, 1995). Een simpele mogelijkheid is stratificatie naar seizoen, e.g. verificatie van extreme winterse omstandigheden beperken tot de wintermaanden. Brooks (2004) beperkt bij de verificatie van tornado warnings het aantal gevallen door alleen die gevallen te beschouwen waarvoor forecasters inschatten dat er 'mogelijkerwijs' een tornado zou kunnen ontstaan, en deze set achteraf aan te vullen met de gemiste events. Mason (1989) probeert op een statistische manier het aantal 'zekere' no-no cases zo groot mogelijk te maken.
Wij hebben gekozen om alleen te kijken naar situaties waarin minstens code Geel gegeven is, i.e. wanneer de kans op meer dan 1 ontlading ergens in de provincie (en dus niet in een SGG) tenminste 30% is. De triviale gevallen waarbij het verwachte onweer 'zeker' niet de schaal van de SGG zal bereiken worden dus niet uitgesloten in onze analyse. De verificatiescores geven hiermee dus een schatting voor hoe goed het verwachtingssysteem discrimineert tussen voorspelde onweergevallen (code Geel) en zware (code oranje / rood) onweer-gevallen. Andere meer objectieve methoden zijn mogelijk, bijvoorbeeld gebaseerd op overschrijding van een bepaalde kansdrempel afkomstig van de verwachtingen uit het KOUW-systeem (Schmeits et al., 2008). Belangrijke predictoren in dit systeem zijn o.a. de verwachte waarden van CAPE en convectieve neerslaghoeveelheid. Iedere stratificatie-methode zal echter van invloed zijn op de uitkomst van de scores.

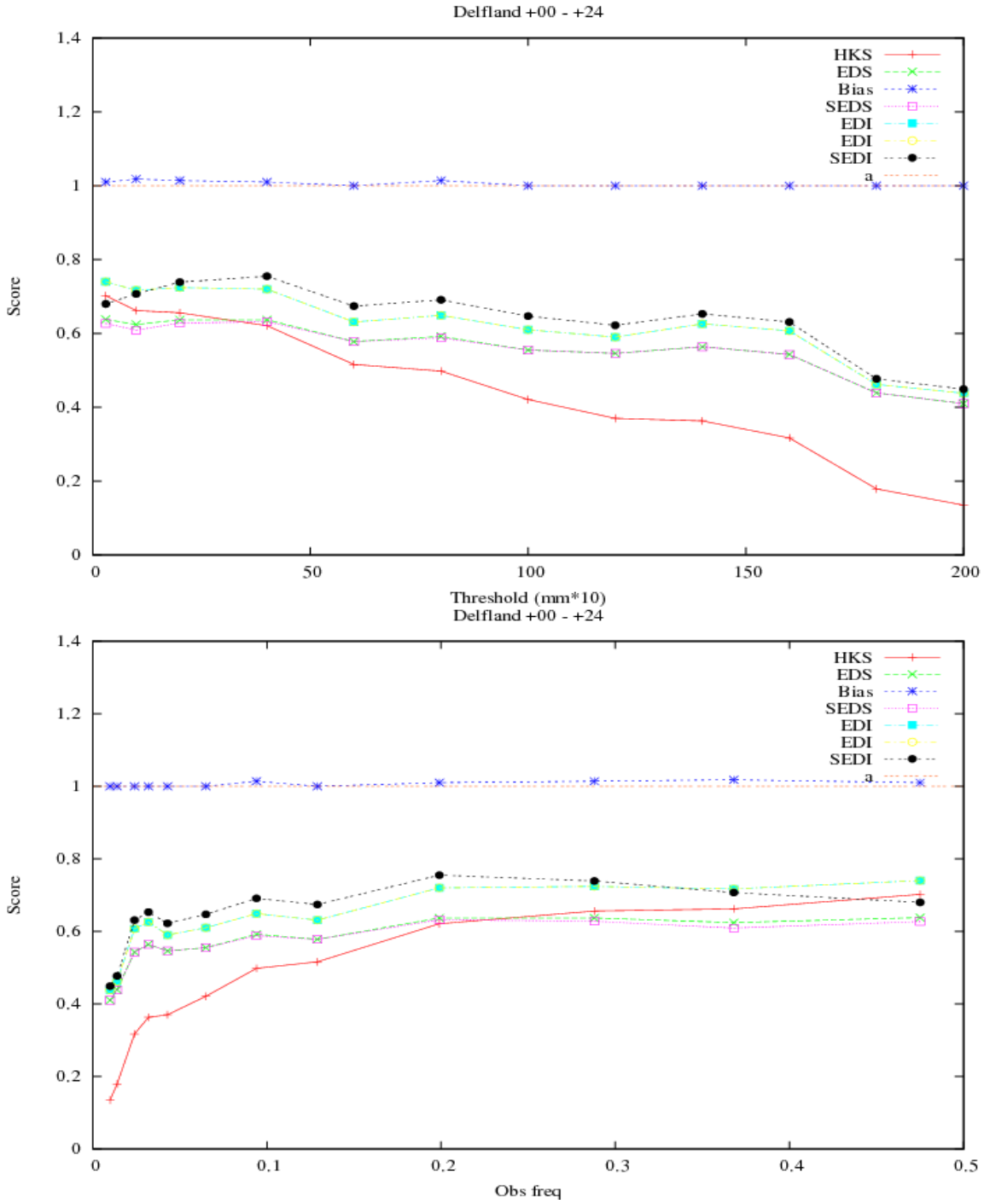


Fig. V.1. Verificatie van gekalibreerde 24-uursneerslagverwachtingen van Hirlam (van 00 tot +24) voor het waterschap Delfland over een periode van ca 5 jaar uitgedrukt in de skill scores HKS, EDS, SEDS, EDI en SEDI, als functie van de waargenomen hoeveelheid (bovenste panel; in tienden mm's) en van de observed frequency (onderste panel). De waarnemingen zijn berekend uit radaraccumulaties voor het betreffende gebied, gekalibreerd met regenmeterwaarnemingen (Holleman, 2007).

VI. Resultaten

In dit hoofdstuk en in de Appendices A1 en A2 worden de resultaten van de verificatie getoond in de opties zoals vermeld in hoofdstuk IV. Alle resultaten hebben betrekking op 2x2 contingentietabellen (Tabel V.1).

De totale periode waarover geverifieerd is loopt van 1 februari 2010 tot 1 juli 2012. Dit komt neer op ruim 21.000 uren. Onze analyse beperkt zich tot de 1978 uren die voldoen aan het criterium dat er minimaal een code Geel voor onweer is afgegeven ergens in Nederland. In Tabel VI.1. staat het aantal uren per provincie waarvoor een code Geel, Oranje of Rood voor onweer gegeven is.

Provincie	GR	FR	DR	FL	OV	GL	UT	NH	ZH	ZE	NB	LB	NL
aantal	1255	1175	1229	1268	1344	1422	1189	1180	1272	1257	1413	1368	1978

Tabel VI.1. Het aantal uren in de periode februari 2010 t/m juni 2012 waarvoor waarschuwingen Geel, Oranje of Rood voor onweer zijn gegeven, uitgesplitst per provincie.

Het accent van de waarschuwingen betrof de oostelijke en zuidoostelijke provincies.
Er is in totaal voor 78 uurvakken voor een of meer provincies code Oranje uitgegeven en voor 15 uurvakken code Rood. Het aantal waargenomen events is 47 voor de optie ‘Midden van SGG’ en 80 voor ‘Corners van SGG’.

Het weglaten van de meest triviale gevallen maakt D in de contingentietabellen een stuk kleiner. Dit heeft geen effect op de zgn. gebruikersscores (POD, FAR, F, CSI) en (dus) ook niet op de bias, maar wel op de skill scores (HKS, SEDS, SEDI, EDI).
Om een significanter beeld te geven worden in onderstaande secties de kwantitatieve resultaten niet voor de afzonderlijke provincies getoond maar voor de vereniging van alle provincies. Hierbij moet opgemerkt worden dat de gevallen geenszins als onafhankelijk beschouwd mogen worden omdat een onweer event meestal meerdere provincies en meerdere uren beslaat. De vereniging van de gevallen van alle provincies wordt in de tabellen weergegeven met ‘**ALL’**. In de grafieken worden de gebruikersscores (m.u.v. de false alarm rate F) wél per provincie getoond.

De resultaten worden vergeleken met een analyse waarbij de provincie-indeling genegeerd wordt (aangeduid als ‘**NL**’). Hiervoor worden de scores berekend op contingentietabellen waarin ‘ergens in Nederland verwacht’ vergeleken wordt met ‘ergens in Nederland voorgekomen’. Deze methode lijkt op de situatie van vóór de regionalisatie van de gevaarlijk weer-systematiek in februari 2010. De analyse bestrijkt heel Nederland inclusief een extra buffer van 25km over de grens (zie Fig. IV.1).

In dit hoofdstuk wordt in de tabellen en grafieken uitgegaan van een drempel van 500 ontladingen per 5 minuten, de drempel waarvoor de verwachtingen gemaakt zijn. Resultaten voor een drempel van 300 ontladingen per 5 minuten zijn in Appendix A1 opgenomen.

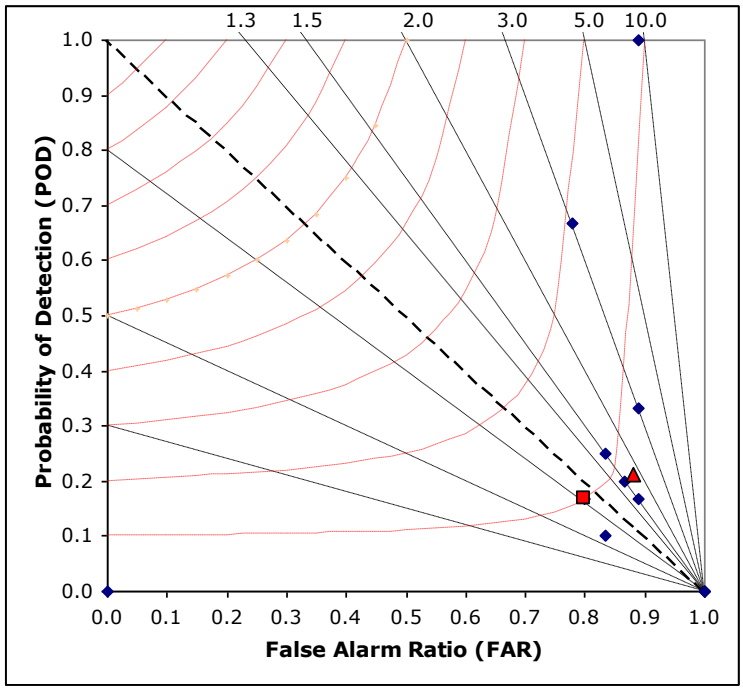
VI. 1. Verificatie onafhankelijk van forecasttijd

In deze paragraaf worden resultaten getoond waarin geen rekening is gehouden met het moment waarop er gewaarschuwd is; alleen het feit of er een waarschuwing is afgegeven of niet telt hier.

A. Midden van SGG

a. Waargenomen onweer (>500 ontladingen per 5 minuten) vs weeralarm (code Rood)

Fig. VI.1 toont de False Alarm Ratio, Probability of Detection, Bias, en de Critical Success Index en bevat tevens een tabel met de waardes van de skill scores.



Individuele provincies

Alle provincies samen

Nederland

ALL		> 500 ontladingen per 5 minuten in 50x50km vak	
		YES	**NO**
Weeralarm (Rood)	**YES**	10	73
	NO	37	15252

NL		> 500 ontladingen per 5 minuten in 50x50km vak	
		YES	**NO**
Weeralarm (Rood)	**YES**	3	12
	NO	15	1948

	POD	FAR	F	CSI	Bias	N	SEDS	HKS	EDI	SEDI
ALL	0.21	0.88	0.00	0.08	1.77	15372	0.501	0.208	0.551	0.565
NL	0.17	0.80	0.01	0.10	0.83	1978	0.476	0.161	0.480	0.492

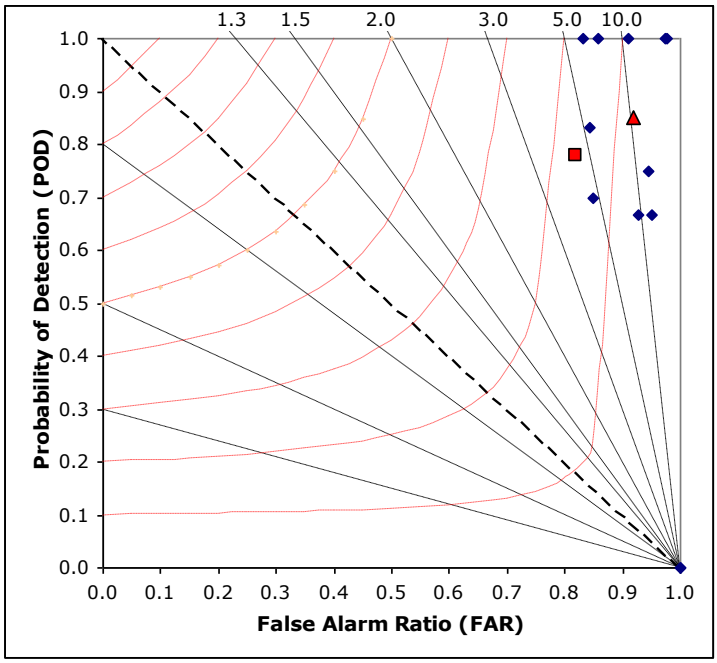
Fig. VI.1. In de POD-FAR plot staat voor de code Rood waarschuwingen de Probability of Detection uitgezet tegen de False Alarm Ratio. Tevens staan de Bias (rechte gestreepte lijnen) en de Critical Success Index (rode gekromde stippellijnen) aangegeven waarbij de waarden van de isolijnen van de bias aan de bovenkant van de figuur staan en die van de CSI bij de snijpunten met de verticale as. De no bias lijn is de dikkere diagonaallijn. Onder de plot staan de contingentietabellen voor alle provincies samen (ALL, links) en voor Nederland (NL, rechts). Onderaan staan de waarden van de verifcatiescores.

Perfecte forecasts vinden we in de linkerbovenhoek van de figuur, waar de POD = CSI = Bias = 1, en de FAR = 0. Punten ver boven (onder) de (gestippelde) diagonaal wijzen op een sterke bias in het systeem: een systeem dat veel vaker (minder vaak) een waarschuwing afgeeft dan de waarnemingen rechtvaardigen.
De scores voor de individuele provincies vertonen zoals verwacht een variabel beeld. Dit wordt veroorzaakt door het lage aantal opgetreden en verwachte gevallen. In een enkele provincie is de POD 1 maar worden er veel false alarms afgegeven. In een ander geval is er voor geen enkele van de zware onweerssituaties gewaarschuwd met code Rood resulterend in een POD van 0.
Voor alle provincies samen, is, met de verificatie volgens 'Midden van SGG', 88% van de waarschuwingen niet terecht (meteorologisch gezien een false alarm). Daarnaast wordt voor slechts ongeveer 1 op de 5 opgetreden weeralarmsituaties gewaarschuwd (POD = 0.21). De overforecasting is iets minder dan een factor 2. Voor NL (zonder regionalisatie) zijn de scores vergelijkbaar, met dien verstande dat de FAR zoals verwacht iets kleiner is (omdat veel locatiefouten niet tellen als false alarm) en de bias is gehalveerd.
De skill scores zijn daarentegen beter voor de geregionaliseerde waarschuwingen dan voor waarschuwingen voor Nederland als geheel. Hoe groot de rol van de 2 maal grotere bias ratio hierin is is echter niet duidelijk. Zoals verwacht liggen de waardes van de HKS dicht bij die van de POD.

b. Waargenomen onweer (>500/5 min) vs waarschuwing voor extreem weer (code Oranje)

De resultaten voor afgegeven code Oranje (Fig VI.2) kennen in vergelijking tot de code Rood waarschuwingen (Fig. VI.1), een veel hogere POD – er wordt voor ongeveer 4 op de 5 onweer events gewaarschuwd – en vergelijkbare FAR-waarden. Dit geldt zowel voor de individuele provincies als voor hun vereniging en voor Nederland als geheel. De combinatie van een hoge POD en hoge FAR wijst op een sterk positieve bias (overforecasting) in het waarschuwingssysteem. Gemiddeld gesproken wordt voor alle provincies samen 10 keer vaker een waarschuwing gegeven dan het aantal daadwerkelijk opgetreden gevaarlijk weersituaties. Voor NL is dit een factor 4.
Verschillende oorzaken kunnen ten grondslag liggen aan deze systematische overschatting. Zo kan het veroorzaakt zijn door een overschatting van de zwaarte of van de impact van het fenomeen. In het eerste geval wordt de kans van overschrijden van een weeralarmdrempel te hoog ingeschat, in het tweede geval wordt bij een te lage kans (lager dan 60%) code Oranje gehesen. Ook kan een *safety first* houding leiden tot overschatting, een houding die voortkomt uit het feit dat gemiste events doorgaans ernstiger consequenties hebben dan false alarms. Tenslotte wordt de overforecasting waarschijnlijk voor een belangrijk deel veroorzaakt doordat forecasters bij onzekerheid over de timing van het fenomeen de waarschuwingscode over een te lange periode afgeven (zie ook hoofdstuk II).

Als we de waarschuwingen zouden verifiëren tegen een lagere drempelwaarde van 300 ontladingen per 5 minuten in een SGG-gebied wordt de bias min of meer gehalveerd. Bovendien vermindert het aantal false alarms met ca 10% (zie Appendix A1).



ALL		>500 ontladingen per 5 minuten in 50x50km vak	
		YES	**NO**
Waarschuwing (Oranje)	**YES**	40	452
	NO	7	14873

NL		>500 ontladingen per 5 minuten in 50x50km vak	
		YES	**NO**
Waarschuwing (Oranje)	**YES**	14	64
	NO	4	1896

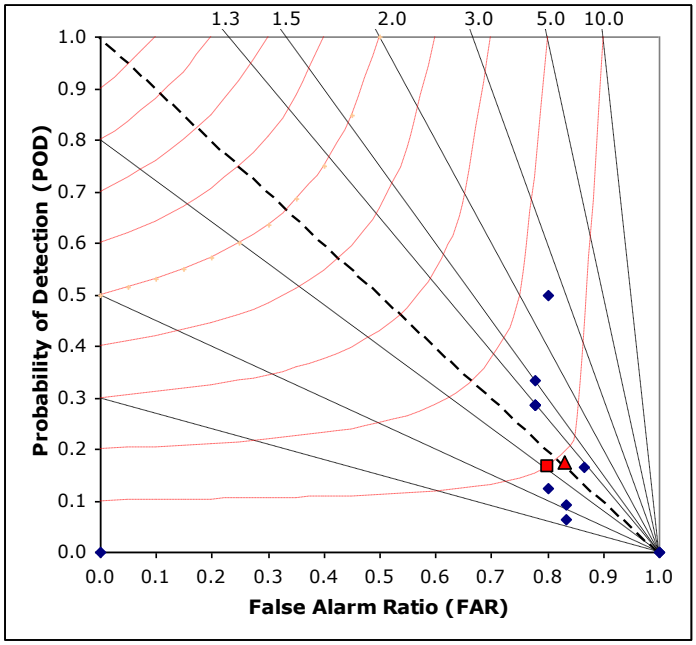
	POD	**FAR**	**F**	**CSI**	**Bias**	**N**	**SEDS**	**HKS**	**EDI**	**SEDI**
ALL	0.85	0.92	0.03	0.08	10.47	15372	0.549	0.822	0.912	0.932
NL	0.78	0.82	0.03	0.17	4.33	1978	0.599	0.745	0.863	0.891

Fig. VI.2. Als Fig. VI.1, maar voor code Oranje.

B. Corners van SGG

Bij het hanteren van de optie ‘Corners van SGG’ verschillen alleen de gecategoriseerde waarnemingen in de contingentietabellen, de waarschuwingen zijn onveranderd. Uiteraard zijn de getallen voor geheel Nederland identiek aan die in sectie A.

a. Waargenomen onweer (>500 ontladingen per 5 minuten) vs weeralarm (code Rood)



Symbool	Uitleg
◆	Individuele provincies
▲	Alle provincies samen
■	Nederland

ALL		> 500 ontladingen per 5 minuten in 50x50km vak	
		YES	NO
Weeralarm (Rood)	YES	14	69
	NO	66	15223

NL		> 500 ontladingen per 5 minuten in 50x50km vak	
		YES	NO
Weeralarm (Rood)	YES	3	12
	NO	15	1948

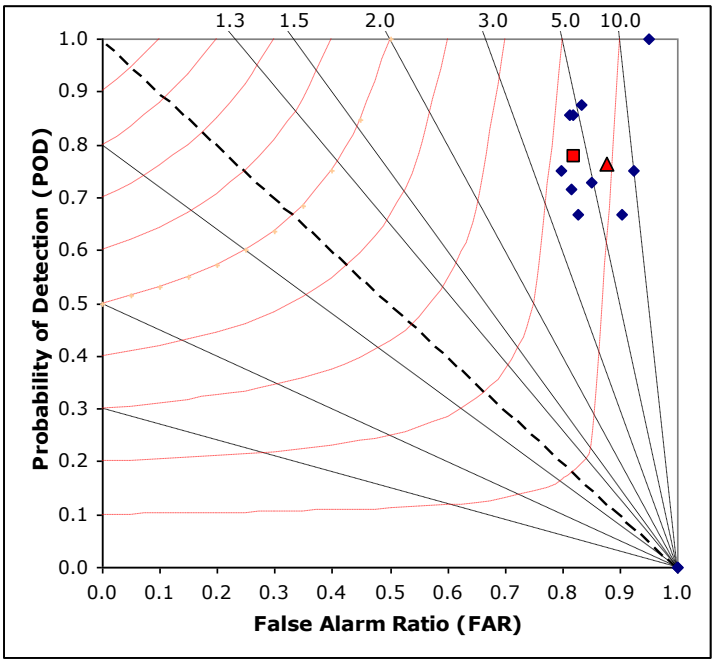
	POD	FAR	F	CSI	Bias	N	SEDS	HKS	EDI	SEDI
ALL	0.18	0.83	0.00	0.09	1.04	15372	0.497	0.170	0.512	0.524
NL	0.17	0.80	0.01	0.10	0.83	1978	0.476	0.161	0.480	0.492

Fig. VI.3. Als Fig. VI.1 voor de code Rood waarschuwingen, maar nu volgens de optie ‘Corners van SGG’.

Vergelijking van bovenstaande verificatie met de code Rood verificatie uitgaande van het midden van het SGG (sectie A) laat zien dat, zoals verwacht, in alle gevallen de positie van

ALL in de figuren dichter bij die van NL komt te liggen. Dit gebeurt doordat zowel de FAR (zoals verwacht) als de POD lager worden (i.e. meer missers en minder false alarms). De bias is vrijwel perfect. De lagere POD wordt veroorzaakt door een aantal situaties waarin het waargenomen gebied (zoals gedefinieerd door 'Corners van SGG') onrealistisch groot wordt. Zie Fig.III.2 voor een illustratie hiervan.

b. Waargenomen onweer (>500/5 min) vs waarschuwing voor extreem weer (code Oranje)



ALL		> 500 ontladingen per 5 minuten in 50x50km vak	
		YES	**NO**
Waarschuwing (Oranje)	**YES**	61	431
	NO	19	14861

NL		> 500 ontladingen per 5 minuten in 50x50km vak	
		YES	**NO**
Waarschuwing (Oranje)	**YES**	14	64
	NO	4	1896

	POD	FAR	F	CSI	Bias	N	SEDS	HKS	EDI	SEDI
ALL	0.76	0.88	0.03	0.12	6.15	15372	0.571	0.734	0.859	0.887
NL	0.78	0.82	0.03	0.17	4.33	1978	0.599	0.745	0.863	0.891

Fig. VI.4. Als Fig. VI.1 voor de waarschuwingen voor extreem weer (code Oranje) voor de optie 'Corners van SGG'.

Vergelijking met de andere optie laat ook hier zien dat (vooral) de POD en de FAR lager zijn (i.e. meer missers en minder false alarms). De bias wordt volgens deze definitie van de waarnemingen beter maar blijft aanzienlijk groter dan 1.

VI. 2. Verificatie als functie van de forecasttijd

In deze sectie worden resultaten voor de optie ‘Midden van SGG’ gegeven als functie van de forecast lead times +0 t/m +6 uur. Resultaten voor ‘Corners van SGG’ staan in Appendix A.2.

Midden van SGG

Code Rood

De resultaten in Fig. VI.5 laten zien dat de scores snel slechter worden met toenemende forecasttijd. De POD is maximaal rond de 20% en de FAR minimaal 50%. Vanaf +3 uur bevinden de punten zich in de rechter benedenhoek. Dit wil zeggen dat in alle gevallen waar 3 uur of langer vooruit een weeralarm afgekondigd werd het criterium niet gehaald is, en dat in alle gevallen waar het criterium wel gehaald is hier niet voor is gewaarschuwd. Dit betekent dat de *waarschuwing*sfunctie wat betreft code rood zeer beperkt is.

Mid ALL Rood 500	A	B	C	D	N	FAR	POD
0 hr	8	16	39	14731	14794	0.67	0.17
1 hr	10	15	37	14732	14794	0.60	0.21
2 hr	3	22	44	14726	14795	0.88	0.06
3 hr	0	24	47	14727	14798	1.00	0.00
4 hr	0	17	47	14740	14804	1.00	0.00
5 hr	0	8	47	14758	14813	1.00	0.00
6 hr	0	5	47	14767	14819	1.00	0.00

Mid NL Rood 500	A	B	C	D	N	FAR	POD
0 hr	2	3	16	1859	1880	0.60	0.11
1 hr	3	3	15	1859	1880	0.50	0.17
2 hr	1	5	17	1857	1880	0.83	0.06
3 hr	0	5	18	1859	1882	1.00	0.00
4 hr	0	3	18	1863	1884	1.00	0.00
5 hr	0	2	18	1868	1888	1.00	0.00
6 hr	0	1	18	1870	1889	1.00	0.00

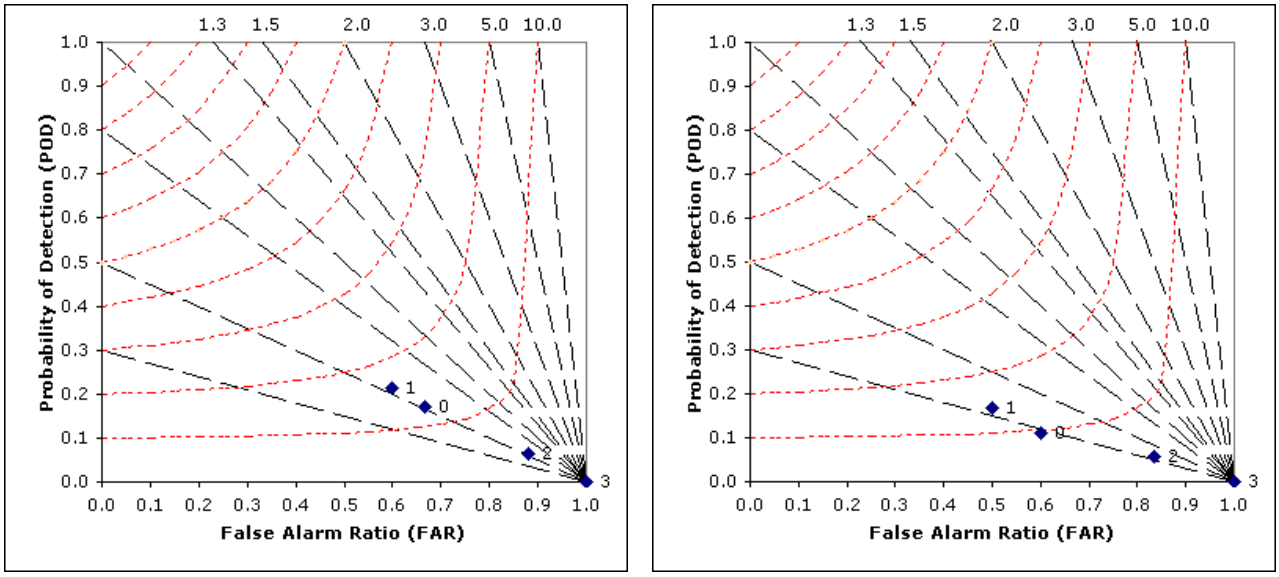


Fig. VI.5. POD-FAR plot voor code Rood waarschuwingen als functie van de forecasttijden vanaf +0 uur tot +3, met boven de plots de bijbehorende elementen van de contingentietabellen voor alle provincies samen (ALL, links) en Nederland (NL, rechts).

Opvallend is dat de +1 uur waarschuwing iets beter scoort dan de +0. Dit verschil wordt veroorzaakt door 1 extra hit (het aantal waarschuwingen wordt in de tabel boven de figuren gegeven door cel A + cel B).

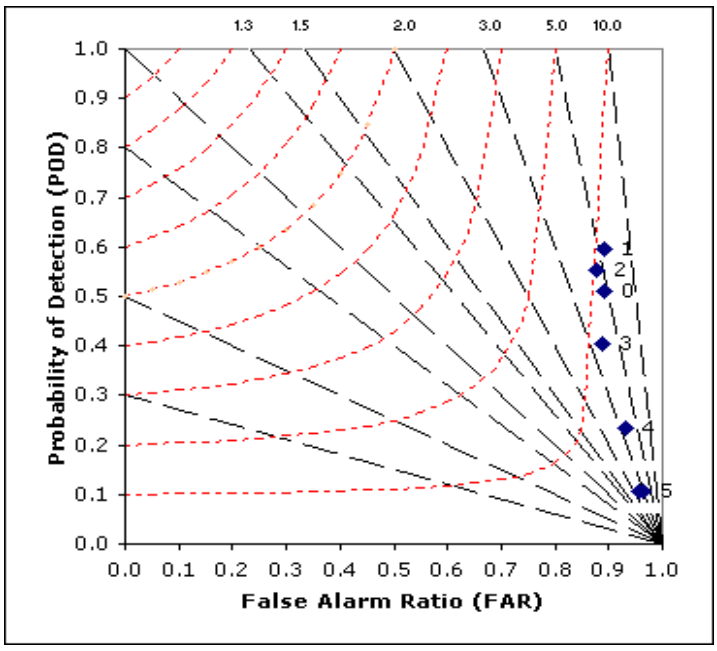
De underforecasting zou verklaard kunnen worden als in veel gevallen de impact bij overschrijding van het onweercriterium als laag ingeschat is.

Code Oranje

Bij de uitgifte van code Oranje treedt juist een grote overforecasting op (Fig. VI.6). Vooral de PODs zijn (dus) veel gunstiger dan bij code Rood. Het aantal hits (cel A) daalt snel met toenemende forcasttijd maar blijft positief tot nog veel verder dan in de figuren getoond is. Tot in ieder geval 12 uur vooruit werd in een aantal gevallen een code Oranje gevolgd door het halen van het weeralarmcriterium.

Mid ALL Oranje 500	A	B	C	D	N	FAR	POD
0 hr	24	198	23	14549	14794	0.89	0.51
1 hr	28	227	19	14520	14794	0.89	0.60
2 hr	26	184	21	14564	14795	0.88	0.55
3 hr	19	149	28	14602	14798	0.89	0.40
4 hr	11	147	36	14610	14804	0.93	0.23
5 hr	5	136	42	14630	14813	0.96	0.11
6 hr	5	112	42	14660	14819	0.96	0.11

Mid NL Oranje 500	A	B	C	D	N	FAR	POD
0 hr	15	33	3	1829	1880	0.69	0.83
1 hr	14	40	4	1822	1880	0.74	0.78
2 hr	10	34	8	1828	1880	0.77	0.56
3 hr	7	25	11	1839	1882	0.78	0.39
4 hr	5	22	13	1844	1884	0.81	0.28
5 hr	3	19	15	1851	1888	0.86	0.17
6 hr	3	15	15	1856	1889	0.83	0.17



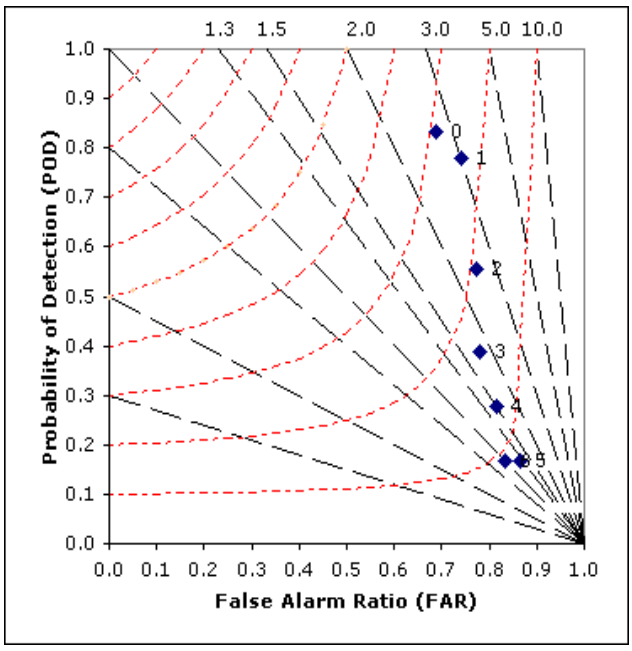


Fig. VI.6. als Fig. VI.5 voor code Oranje.

VII. Samenvatting en conclusies

De kwaliteit van de gevaarlijk weer verwachtingen is van essentieel belang voor het KNMI. Om deze kwaliteit zichtbaar te maken en om eventuele tekortkomingen en mogelijke verbeteringen te identificeren is een evaluatie en verificatie nodig, niet alleen van het 'eindproduct' maar ook van alle tussenstappen die leiden tot het al of niet waarschuwen.
De allereerste stap in deze keten is de meteorologische beoordeling door een gezamenlijk oordeel van de meteorologen in de weerdienst. In de (eventuele) volgende stappen wordt de inschatting van mogelijke 'impact' van de mogelijke gevaarlijk weer situatie steeds groter en is uiteindelijke bepalend voor het al of niet uitgeven van een waarschuwing voor extreem weer of een weeralarm (zie Fig. I.1).

In februari 2010 is de gevaarlijk weer/weeralarm systematiek grondig gewijzigd. Een van de belangrijkste redenen hiervoor was het (te) grote aantal false alarms. Voor een vast aantal afgesproken predictands wordt vanaf dat moment voor ieder uur en voor iedere provincie het waarschuwingsniveau uitgegeven in de vorm van een kleurcodering lopend van Groen (geen waarschuwing) via Geel, Oranje tot Rood (weeralarm). Voor de 2 hoogste waarschuwings-niveaus is het KNMI niet meer alleen verantwoordelijk maar een weeralarmteam waarin verschillende maatschappelijke organisaties een stem hebben. In 2015 is de systematiek vervangen door een nieuwe systematiek, beschreven in "Herijking Waarschuwings-systematiek 2015" (KNMI rapport, juni 2015).

In dit rapport, dat betrekking heeft op de waarschuwingssystematiek die gehanteerd werd van februari 2010 tot september 2015, is gekeken naar de uitgegeven waarschuwingen in de kleurcodes Rood en Oranje zoals ze op Internet te zien zijn voor het 'brede publiek'. Er is vooralsnog alleen gekeken naar het onweercriterium: een aantal ontladingen van minimaal 500 per 5 minuten in een standaard gebiedsgrootte (SGG) van 50 x 50 km. De periode betreft februari 2010 tot juli 2012 en bevat een relatief klein aantal events: het betrof slechts 10 onweer events die voldeden aan het weeralarmcriterium, maar die zich wel over meerdere provincies en uurvakken uitstrekten. Hoewel voor de uitgifte van de codes Oranje en Rood de verwachte impact leidend is, is impact als zodanig niet meegenomen in de verificatie. Deze studie geeft daarom uitsluitend inzicht in de relatie (coïncidentie) tussen de al of niet uitgegeven waarschuwing en de mate van bliksemactiviteit.

Uit deze studie kunnen geen conclusies getrokken worden over alle deelprocessen in de keten van de gevaarlijk weer systematiek. Hiertoe dienen alle stappen apart geëvalueerd en geverifieerd te worden. Een eerste aanzet tot verificatie van het meteorologische onderdeel is gedaan in Kok et al (2011b). Voor de andere stappen in de keten is niet eerder een systematische evaluatie gedaan.

De reden waarom in deze verificatie begonnen is met onweer is omdat hiervan redelijk dekkende waarnemingen bekend zijn in ruimte en tijd, afkomstig van het FLITS detectiesysteem, waardoor met redelijke zekerheid bepaald kan worden of het waarschuwingscriterium gehaald is of niet. Voor de meeste andere predictands in de gevaarlijk weer systematiek is dit zeker niet het geval. Hiertoe dient een andere benadering

gekozen te worden, zoals beschreven in Kok (2005). De in dit rapport gekozen methode is daarom niet direct overdraagbaar naar de meeste andere predictands. In de in Kok (2005) voorgestelde aanpak wordt de onzekerheid over het al of niet optreden van een event expliciet meegenomen in de verificatie en uitgedrukt in de vorm van een kans, in plaats van als een 'ja' of een 'neen' uitspraak. Deze kansen kunnen vervolgens op een standaard manier vergeleken worden met de (categorische) waarschuwingen (e.g. Kok et al., 2011b).

De categorische (wel/niet) waarschuwingen zijn hier geverifieerd tegen categorische waarnemingen, i.e. ja/nee opgetreden. De beoordeling of iets wel of niet is opgetreden kan op meerdere criteria gebaseerd worden. Zo kunnen we een waargenomen event vaststellen als het centrum van de onweersactiviteit *in* de betreffende provincie ligt, maar ook als de onweersactiviteit een provincie '*raakt*', i.e. als een deel van de provincie onderdeel uitmaakt van het gebied van 50km x 50km waarin het criterium overschreden is. De tweede definitie is veel minder streng dan de eerste en gaf een lagere FAR (zoals verwacht) maar in de meeste gevallen ook een lagere POD.

De uitgegeven waarschuwingen zijn op verschillende manieren op hun kwaliteit beoordeeld. Allereerst is dit gebeurd in het detail waarop ze uitgegeven zijn, namelijk per uurvak per provincie. Verificatie laat zien dat weeralarmen vanaf 3 uur vooruit *meteorologisch gezien* alleen maar misses en false alarms opleveren. Code Oranje geeft daarentegen weliswaar ook veel false alarms maar relatief veel minder gemiste events. Dit is het gevolg van een grote mate van overforecasting. Dit zou een indicatie kunnen zijn dat de impact overschat wordt of dat er al veel impact optreedt (en verwacht wordt) bij een lagere meteorologische drempel. Daarnaast valt op dat in slechts 20% van de weeralarmsituaties een weeralarm afgegeven is in het uur waarin het criterium overschreden wordt of al is overschreden. Deze grote 'onder-waarschuwing' kan veroorzaakt zijn door een lage-impact inschatting.

Opvallend is dat tot (tenminste) 12 uur vooruit de code Oranje verwachtingen nog steeds enige 'skill' vertonen. Tot die termijn kan er dus wel degelijk nog sprake zijn van een nuttige waarschuwingsfunctie. Verificatie tegen een iets lagere drempel van 300 ipv 500 ontladingen per 5 minuten in het SGG doet het aantal extremen verdubbelen en het aantal false alarms afnemen met zo'n 10%. Deze beide bevindingen zijn indicaties dat er meteorologisch gezien veel potentie zit om te komen tot extreem weer verwachtingen die 'de maatschappij' veel beter zouden kunnen bedienen. De verificatie tegen de lagere drempel toonde daarentegen ook aan dat het aantal near-misses groot was.

Tenslotte is de geregionaliseerde systematiek vergeleken met de systematiek van vóór februari 2010 door de regionale waarschuwingen weer te aggregeren tot landelijke. De overforecasting bij code Oranje neemt dan duidelijk af, maar is nog steeds ongeveer een factor 4. Een belangrijk kritiekpunt van de vroegere 'landelijke' weeralarmen was dat een groot deel van Nederland moest 'lijden' onder de acties die volgden op een waarschuwing die eigenlijk maar voor een deel van het land bedoeld was. Uit de verificatie blijkt dat hoewel de False Alarm Ratio (de fractie false alarms op het totaal van uitgegeven waarschuwingen) uiteraard groter is geworden (door te kijken per provincie), het aantal mensen dat 'false' gealarmeerd wordt zeker is gehalveerd. In die zin heeft de in 2010 ingevoerde systematiek aan haar doel voldaan.

VIII. Aanbevelingen en discussie

Op basis van deze verificatie en gesprekken met meteorologen kan een aantal aanbevelingen gedaan worden over de waarschuwingssystematiek die operationeel was van februari 2010 t/m september 2015. In deze paragraaf beperken we ons tot de in deze studie opgevallen zaken, die dus voornamelijk betrekking hebben op onweer.

De aanbevelingen kunnen globaal verdeeld worden in 4 categorieën:

1. **Logistiek**
   - De beschikbaarheid van data in het geschikte formaat was niet optimaal. Delen staan in UTC en andere in lokale tijd. Schoning, stroomlijning en een betere beschrijving zijn dringend gewenst.
   - Hetzelfde geldt voor de archivering. Veel data werden aangetroffen in 'persoonlijke' archieven waarvan de borging niet gegarandeerd is. Een centrale archivering van deze data en tools om deze makkelijk te ontsluiten is aan te bevelen.

2. **Meteorologisch criterium**
   - De definitie van de SGG kan worden heroverwogen. De meest gebruikte module voor het signaleren van het aantal ontladingen is een vierkant van 50x50km dat in 'oost-west richting' georiënteerd is. Hierdoor kunnen verschillende beoordelingen ontstaan afhankelijk van de oriëntatie van het weerfenomeen. Dit kan opgelost worden door willekeurige oriëntatie van het vierkante gebied toe te staan. Dit betekent dat vierkanten van 50x50 km gezocht worden waarin aan het criterium voldaan is in een gebied dat een factor $\pi/2$ groter is dan de SGG. Ook vervanging door een cirkelvormig gebied of een vormloos aaneengesloten oppervlak kan veel onduidelijkheden voorkomen.
   - Vastgelegd moet worden hoe om te gaan met de SGG bij de landsgrenzen, land-zee overgangen, Waddeneilanden en IJsselmeer.
   - Een definitie van het begrip 'coherente band' is afwezig. In de beoordeling van het al of niet opgetreden of verwachte onweercriterium speelt de coherente band geen rol. Er wordt slechts gekeken naar het aantal ontladingen in de SGG.
   - Het onweercriterium bevat de overbodige toevoeging 'al dan niet met hagel'. Aan deze voorwaarde wordt altijd voldaan. De meteorologische criteria dienen goed gedefinieerde fenomenen te zijn, die zoveel mogelijk *mece* zijn: mutually exclusive and collectively exhaustive (wederzijds uitsluitend en gezamenlijk uitputtend). I.e. er mag geen misverstand bestaan of een bepaald fenomeen aan het criterium voldoet of niet. (Er mag uiteraard wel onduidelijkheid bestaan over de verwachting of de waarneming van het fenomeen). Eventuele toevoegingen van verduidelijkende teksten, zoals 'al of niet met hagel' kunnen in de tekst naar buiten gecommuniceerd worden maar horen niet in de definitie te staan.

3. **Systematiek**
   - De hele gevaarlijk weer systematiek is er op gericht om zo *snel* en zo *goed* mogelijk de 'maatschappij' te waarschuwen voor 'high impact weather'. Er moet

dus actief gezocht worden naar verbeteringen en tekortkomingen. Het wordt aanbevolen om iedere stap in het proces te evalueren en verifiëren, incl. de inbreng van derden. Hiertoe is documentatie en archivering van al deze stappen cruciaal.

- De tijd tussen het besluit tot afgifte van een waarschuwing code Oranje of Rood en het tijdstp waarop die kleur op Internet komt is voor het fenomeen Onweer in enkele gevallen (veel) meer dan 20 minuten. Dit is erg lang voor een waarschuwing voor extreem weer of voor een weeralarm. Dit zou wellicht verkort kunnen worden door de impact assessment in een vroeger stadium te doen (bijvoorbeeld al bij een lagere kans), en de waarschuwing uit te laten gaan zodra de kansdrempel voor het optreden van het betreffende fenomeen overschreden wordt.
- Het streven om het aantal false alarms te beperken leidt onvermijdelijk tot een toename van het aantal misses (e.g. Halsey, 1995; Kok, 2000) en / of tot het zeer laat waarschuwen. “Pas wanneer met tenminste 90% zekerheid kan worden gezegd waar en wanneer het fenomeen zich voordoet wordt het weeralarm uitgegeven” (Een vernieuwd weeralarm, KNMI 2010). Het komt dan ook vrijwel niet voor dat er meer dan 3 uur vantevoren een code rood wordt uitgegeven. Veel andere waarschuwingssystemen waarschuwen bij een veel kleinere zekerheid teneinde juist het aantal misses te beperken. Misses zijn voor de meeste gebruikers namelijk veel kostbaarder dan false alarms. In veel van de Met Office defence warnings wordt een waarschuwing met een lead time van minder dan 3 uur in de verificatie geïnterpreteerd als een miss omdat de response tijd voor de gebruikers te kort is (Stephenson et al., 2010). De waarschuwingen van het Amerikaanse Storm Prediction Center voor een aantal gevaarlijk weer fenomenen (tornado’s, storms) worden al uitgegeven bij een kansdrempel, afhankelijk van element en forecasttijd, variërend van 10 tot 40% (ww.spc.noaa.gov). In het waarschuwingssysteem voor grote hoeveelheden neerslag t.b.v. de waterschappen in Nederland (Kok et al., 2011a), operationeel sinds 2003, variëren de door de waterschappen zelf gekozen kansdrempels van 20 tot 35%. Dit soort waardes is veel meer in overeenstemming met de drempels waarop er maximaal (economisch) effect van het treffen van voorzorgsmaatregelen te verwachten is (zie volgende bullit) en geeft gebruikers bovendien meer tijd.
- Het feit dat er in de waarschuwingssystematiek alleen in categorische zin gewaarschuwd kan worden beperkt zijn *waarde* voor grote groepen gebruikers. Die waarde (value) wordt vaak afgemeten aan de mate waarin waarschuwingen bijdragen aan het decision-making process (beslissingsondersteuning) en wordt meestal aangetoond in een zgn. cost-loss ratio context waarin de kosten (costs C) ter voorkoming of vermindering van schade vergeleken worden met de verliezen (losses L) die optreden door het (extreme) weer als er geen voorzorgsmaatregelen getroffen zijn (e.g. Bilham, 1922; Bijvoet en Bleeker, 1951; Katz and Murphy, 1997; Richardson, 2000). De C-L ratio heeft voor verschillende gebruikers (gebruikersproblemen) verschillende ‘waardes’ (maar meestal tussen 0 en 0.2; Katz and Murphy, 1997). Kansverwachtingen hebben potentieel veel meer waarde voor gebruikers dan categorische of deterministische verwachtingen (Brier, 1950; Thompson, 1952; Murphy, 1977; 1991; 1993). Ter vergroting van de waarde van

de waarschuwingen is het dan ook aan te bevelen om bijv. de kans-inschattingen die gemaakt worden in stap 1 in het proces (zie Fig. I.1) en die de basis vormen van de (meteo) verwachting naar de gebruikers te communiceren. In een studie van Joslyn and LeClerc (2012) werd aangetoond dat betrouwbare informatie over de onzekerheid van de verwachtingen niet alleen leidde tot betere besluiten maar dat ook het vertrouwen in de provider van de verwachtingen groeide. LeClerc and Joslyn (2015) toonden bovendien aan dat communicatie van de onzekerheid sterk de voorkeur verdient boven het verlagen van het aantal false alarms.

4. **Verificatie**

- Wat betreft de metriek die gebruikt wordt bij de verificatie van de (categorische) verwachtingen wordt aanbevolen om, naast een aantal 'gebruikersscores', bij voorkeur de POD (Probability of Detection) en FAR (False Alarm Ratio), een of meerdere zgn. *process oriented scores* (Göber, 2009), of skill scores, te berekenen die onafhankelijk zijn van hoe vaak het te onderzoeken fenomeen voorkomt. Op die manier is de waarde (het getal) van de score meer een maat voor de "skill" van de verwachtingen en minder een maat die bepaald wordt door de 'toevallige' observed frequency van het fenomeen in de bekeken periode. Hierdoor zijn trends in de skill van de verwachtingen beter te monitoren. De Extremal Dependence Index (EDI) en de Symmetric Extremal Dependence Index (SEDI) lijken hiervoor een betere keuze dan de tot nu toe veel gebruikte Hanssen Kuipers Score (HKS), omdat ze naast bovengenoemde eigenschap minder afhankelijk zijn van de (fysieke) drempelwaarde van de predictand en niet naar nul gaan bij rare events (Ferro and Stephenson, 2011). Deze scores zijn echter in meer of mindere mate gevoelig voor bias en moeten dan ook vooraf gekalibreerd worden en altijd in combinatie met de bias gepresenteerd worden. Meer onderzoek en ervaring is nog nodig.

  Een voor de hand liggende manier om de verwachtingen te kalibreren is door gebruik te maken van de kansschattingen die de basis (de eerste stap) vormen in de gevaarlijk weer systematiek. Door een geschikte keuze van een "drempelkans" kunnen deze probabilistische verwachtingen getransformeerd worden naar unbiased categorische verwachtingen. Overigens kunnen bij goede archivering met terugwerkende kracht allerlei (eventueel nieuwe) scores uitgerekend worden en hoeft de keuze van de metriek niet op voorhand vastgelegd te worden.
- Bij categorische rare event forecasting is stratificatie van de data noodzakelijk om er voor te zorgen dat de scores (m.n. de skill scores) niet gedomineerd worden door de talrijke gemakkelijk te verwachten 'niks aan de hand' gevallen. In dit rapport hebben we gekozen te stratificeren naar moeilijkheidsgraad, door gebruik te maken van de door de meteorologen verwachte code Geel situaties. Het verdient aanbeveling om een meer objectieve stratificatie te ontwikkelen die, analoog aan het KOUW-systeem (Schmeits et al., 2008), gebaseerd is op een statistische (probabilistische) relatie tussen enerzijds predictoren afkomstig van NWP model output (CAPE, convectieve neerslag, ice content, etc) aangevuld met andere predictoren, en anderzijds de overschrijding van een vastgesteld maar laag aantal waargenomen ontladingen in gebieden ter grootte van de SGG. Zowel dit aantal als de (drempel)kans van overschrijding dienen zo gekozen te worden dat

het aantal missed events in het weggelaten deel van de data erg klein is. Bij monitoring van de skill over meerdere jaren dient de stratificatie over de hele periode uiteraard hetzelfde te zijn.

- Verificatieresultaten van categorische verwachtingen van extreem weer fenomenen worden meestal als teleurstellend ervaren omdat ze geen weerspiegeling zijn van het vaak positieve oordeel die door gebruikers hebben over de waarschuwingen. Dit is inherent aan het feit dat er in het algemeen geen ruimte is voor onderscheid tussen bijv. flagrante missers en de 'net aan' missers. In het geval van onweer kunnen een paar ontladingen meer of minder immers bepalend zijn of een criterium gehaald is of niet. Het verdient dan ook aanbeveling om naast de strikte grenzen in de meteorologische criteria ook te kijken naar iets minder extreme drempelwaardes, temeer omdat de zgn. *close calls* en *near misses* ook sterk meetellen in het vertrouwen van het publiek in de uitgegeven waarschuwingen (Barnes et al., 2007). In dit rapport is dit gedaan door ook de scores te berekenen voor een grens van 300 i.p.v. 500 ontladingen. Sharpe (2016) heeft een methode ontwikkeld waarin naast een (iets) lagere intensiteit van het fenomeen ook de bijna hits qua locatie en qua timing meegewogen worden.
- Verificatie heeft i.h.a. als belangrijk doel terugkoppeling te geven in de vorm van kwaliteitsmaten teneinde de kwaliteit te bewaken en te verbeteren. Omdat deze studie de uitgegeven waarschuwingen beschouwt waarin een inschatting van de impact een grote rol speelt, en niet de meteorologische kansinschattingen, zijn slechts beperkte conclusies te trekken. Hiervoor is een evaluatie nodig van de kansinschattingen van zowel vóór als na consultatie met de providers. Hiertoe dienen de kansschattingsformulieren consequent ingevuld en gearchiveerd te worden. Dit is ook noodzakelijk om de kansen eventueel te gebruiken voor kalibratie. Om de terugkoppeling naar meteorologen te bevorderen zou daarnaast gedacht kunnen worden om ook voor een aantal sub-extremen, die per definitie veel vaker voorkomen, kansschattingsformulieren in te gaan vullen. Verificatie hiervan kan helpen om de grote bias in de kansschattingen voor de weeralarm-criteria (Kok, 2011b) te corrigeren.

Dankbetuiging

Speciaal voor deze pilot is de onweermodule van Rudolf van Westrhenen die op historische data het maximale aantal bliksemontladingen telt, uitgebreid met een optie die ook locale maxima berekent (i.p.v. landelijke). Gerrit Burgers, Maurice Schmeits, Rutger Boonstra, Adri Huiskamp, Wim de Rooy, Jan Barkmeijer, Henk van den Brink en Bart van den Hurk worden bedankt voor het kritisch doorlezen van eerdere versies van dit rapport. De figuren in hoofdstuk III zijn afkomstig van Rudolf van Westrhenen.

Referenties

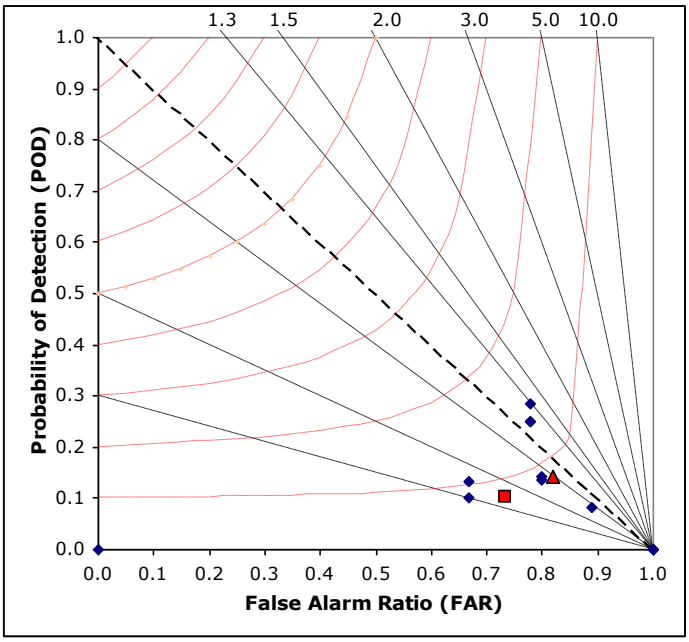
- Barnes, L. R., M. H. Hayden, D. M. Schultz, and C. Benight, 2007. False alarms and close calls: a conceptual model of warning accuracy. *Wea. Forecasting*, **22**, 1140-1147.
- Bijvoet, H. and W. Bleeker, 1951. The value of weather forecasts. Weather **6**, 36-39.
- Bilham. E. G., 1922. A problem in economics. *Nature*, No. 2733, **109**, 341-342.
- Brier, G. W., 1950. Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev*., **78**, 1-3.
- Brooks, H. E., 2004. Tornado-warning performance in the past and future. *Bull. Amer. Meteor. Soc.*, **85**, 837-843.
- ECMWF-TAC report, 2010. Report to the Technical Advisory Committee from the Subgroup on Verification Measures, 2009-2010 (available from www.ecmwf.int).
- Ferro, C. A. T., and D. B. Stephenson, 2011. Extremal dependence indices: improved verification measures for deterministic forecasts of rare binary events. *Wea. Forecasting*, **26**, 699-713.
- Ghelli, A., and C. Primo, 2009. On the use of the extreme dependency score to investigate the performance of an NWP model for rare events. *Meteor. Appl*., **16**, 537-544.
- Gandin, L. S., and A. H. Murphy, 1992. Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.*, **120**, 361-370.
- Göber, M., 2009. Tutorial on warning verification. *International Verification Methods Workshop.* (beschikbaar via http://space.fmi.fi/Verification2009/)
- Halsey, N. G. J., 1995. setting verification targets for minimum road temperature forecasts. *Meteorol. Appl*, **2**, 193-197.
- Hanssen, A. W. and J. A. Kuipers (1965). On the relationship between the frequency of rain and various meteorological parameters. *Mededelingen en verhandelingen* **81**. KNMI publ. 68pp.
- Hogan, R. J., E. J. O'Connor, and A. J. Ilingworth, 2009. Verification of cloud-fraction forecasts. *Quart. J. Roy. Meteor. Soc.*, **135,** 1494-1511.
- Holleman, I., 2007. Bias adjustment and long-term verification of radar-based precipitation estimates. *Meteorol. Appl*, **14**, 195-203.
- Joslyn, S, and J. E. LeClerc, 2012. Uncertainty forecasts improve weather related decisions and attenuate the effects of forecast error. *J. Exp. Psychol. Appl*., **18**,126/140.
- Katz, R. W. and A. H. Murphy (eds), 1997. Economic value of weather and climate forecasts. Combridge University Press, 222pp.

- KNMI 2010. Een vernieuwd weeralarm. Aanpassing weeralarmcriteria en -systematiek per 1 februari 2010. KNMI publicatie.
- KNMI 2010. 'Uitgifteproces Waarschuwingen en weeralarmen' (12 januari 2010).
- Kok, C. J., 2000. On the behaviour of a few popular verification scores in yes/no forecasting. Scientific Report WR-2000-04. 73pp.
- Kok, Kees, 2005. Naar een andere weeralarmsystematiek. KNMI Memorandum WM 05-02. Mei 2005.
- Kok, C. J., B. G. J. Wichers Schreur and D. H. P. Vogelezang, 2011a. Meteorological support for anticiparory water management. *Atmos. Res*. **100**, 285-295.
- Kok, Kees, Frits Koek and Ben Wichers Schreur, 2011b. First assessment of extreme weather warnings in Holland. EMS-poster. Berlijn, September 2011.
- LeClerc, J., and S. Joslyn, 2015. The Cry Wolf effect and weather-related decision making. *Risk Analysis*, **35**, 385-395.
- Mason, I., 1989. Dependence of the Critical Success Index on sample climate and threshold probability. *Austr. Met. Mag.*, **37**, 75-81.
- Mureau, R., 2005. Verificatie van weeralarmen 1999-2005 (te zien in Kok, 2005).
- Murphy, A. H., 1977. The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Mon. Wea. Rev*., **105**, 803-816.
- Murphy, A. H., 1991. Probabilities, odds, and forecasts of rare events. *Wea. Forecasting*, **6**, 302-307.
- Murphy, A. H., 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281-293.
- Murphy, A. H., 1995. A coherent method of stratification within a general framework for forecast verification. *Mon. Wea. Rev*., **123**, 1582-1588.
- Noteboom, S. 2006. Processing, validatie, en analyse van bliksemdata uit het SAFIR/FLITS systeem. KNMI Intern Rapport IR 2006-1.
- Richardson, D., 2000. Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc*., **126**, 649-667.
- Roebber, P. J., 2009. Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601-608.
- Schmeits, Maurice J., Kees J. Kok, Daan H. P. Vogelezang, and Rudolf M. van Westrhenen, 2008. Probabilistic forecasts of (severe) thunderstorms for the purpose of issuing a weather alarm in the Netherlands. *Wea. Forecasting*, **23**, 1253-1267.
- Sharpe, M. A., 2016. A flexible approach to the objective verification of warnings. *Meteorol. Appl*., **23**, 65-75.
- Stephenson, D. B., B. Casati, C. A. T. Ferro, and C. A. Wilson, 2008. The extreme dependency score: a non-vanishing measure for forecasts of rare events, *Meteorol. Appl.*, **15**, 41-50.
- Stephenson, D. B., I. T. Joliffe, C. A. T. Ferro, C. A. Wilson, M. Sharpe, M. Mittermaier, and T. D. Hewson, 2010. White paper review on the verification of warnings. UK Met Office Techn. Report No. 546.
- Thompson, J. C., 1952. On the operational deficiences in categorical weather forecasts. *Bull. Amer. Meteor. Soc*., **33**, 223-226.
- Wessels, H. R. A., 1998. Evaluation of a radio interferometry lightning positioning system. KNMI Scientific Rep. WR-98-04, De Bilt Netherlands, 26pp.

APPENDIX A1. Verificatie onafhankelijk van forecasttijd bij een drempel van 300 ontladingen per 5 minuten

A. Midden van SGG

a. Waargenomen onweer (>300 ontladingen per 5 minuten) vs weeralarm (code Rood)



Symbool	Uitleg
◆	Individuele provincies
▲	Alle provincies samen
■	Nederland

ALL		>300 ontladingen per 5 minuten in 50x50km vak	
		YES	**NO**
Weeralarm (Rood)	**YES**	15	68
	NO	90	15199

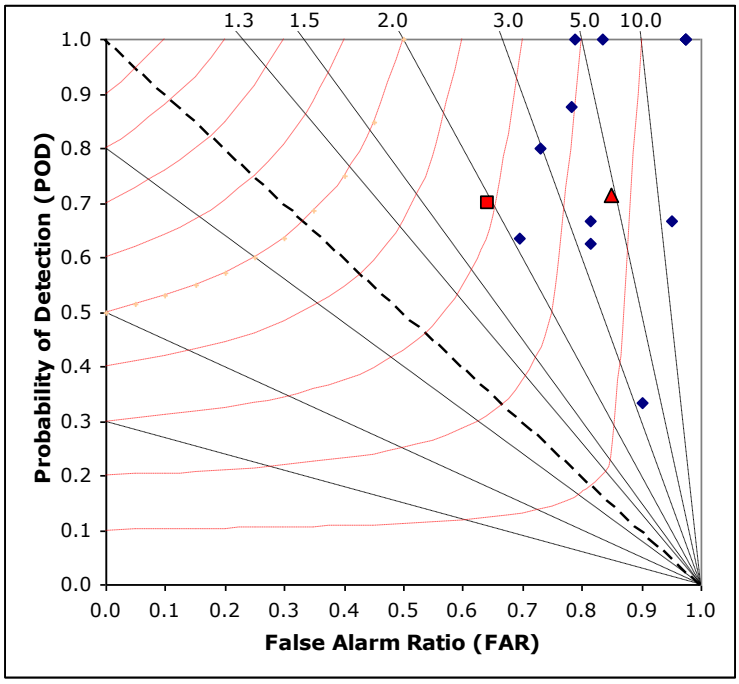
NL		>300 ontladingen per 5 minuten in 50x50km vak	
		YES	**NO**
Weeralarm (Rood)	**YES**	4	11
	NO	36	1927

	POD	FAR	F	CSI	Bias	A	B	C	D	N	SEDS	HKS	EDI	SEDI
ALL	0.14	0.82	0.00	0.09	0.79	15	68	90	15199	15372	0.473	0.138	0.471	0.481
NL	0.10	0.73	0.01	0.08	0.38	4	11	36	1927	1978	0.416	0.094	0.384	0.391

Fig. A1.1. In de POD-FAR plot staat voor de code Rood waarschuwingen de Probability of Detection uitgezet tegen de False Alarm Ratio. Tevens staan de Bias (rechte gestreepte lijnen) en de Critical Success Index (rode gekromde stippellijnen) aangegeven waarbij de waarden van de isolijnen van de bias aan de bovenkant van de figuur staan en die van de CSI bij de snijpunten met de verticale as. De no bias lijn is de dikkere diagonaallijn. Onder de plot staan de contingentietabellen voor alle provincies samen (ALL, links) en voor Nederland (NL, rechts). Onderaan staan de waarden van de verifcatiescores.

De resultaten voor code Rood geverifieerd tegen 300 of meer ontladingen per 5 minuten in een gebied ter grootte van de SGG staan in Fig. A1.1. Door dit lagere waarneemcriterium wordt het aantal uren dat aan het criterium voldoet ruim verdubbeld. Ten opzichte van het 500 ontladingen per 5 minuten criterium wordt de FAR lager zoals verwacht. Een aantal gevallen met tussen de 300 en 500 ontladingen wordt nu als hit i.p.v. false alarm aangemerkt. De POD daalt echter ook. Blijkbaar wordt een groot percentage van de 'nieuwe' observed cases gemist.

b. Waargenomen onweer (>300/5 min) vs waarschuwing voor extreem weer (code Oranje)



ALL		>300 ontladingen per 5 minuten in 50x50km vak	
		YES	**NO**
Waarschuwing (Oranje)	**YES**	75	417
	NO	30	14850

NL		>300 ontladingen per 5 minuten in 50x50km vak	
		YES	**NO**
Waarschuwing (Oranje)	**YES**	28	50
	NO	12	1888

	POD	FAR	F	CSI	Bias	A	B	C	D	N	SEDS	HKS	EDI	SEDI
ALL	0.71	0.85	0.03	0.14	4.69	75	417	30	14850	15372	0.581	0.687	0.829	0.860
NL	0.70	0.64	0.03	0.31	1.95	28	50	12	1888	1978	0.672	0.674	0.822	0.854

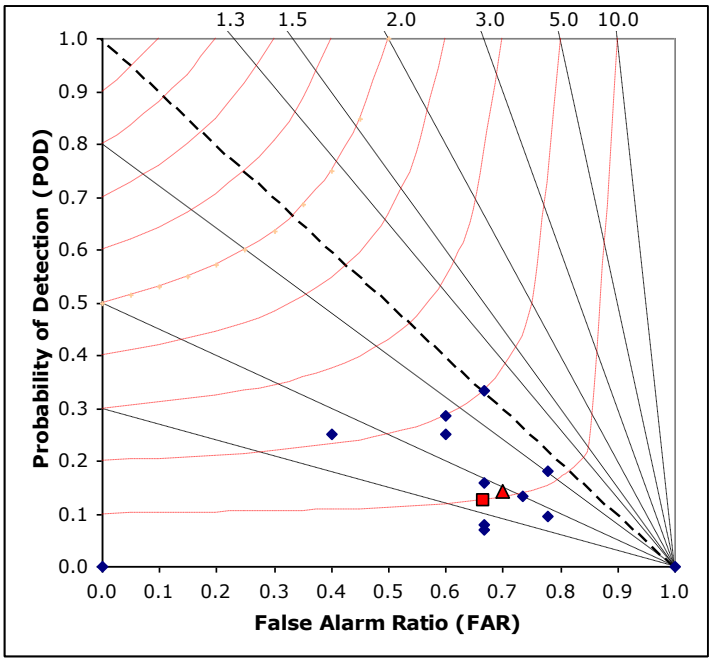
Fig. A1.2. Als Fig. A1.1 maar dan voor code Oranje.

De conclusies zijn analoog aan die uit bovenstaande paragraaf voor code Rood.

B. Corners van SGG

a. Waargenomen onweer (>300 ontladingen per 5 minuten) vs weeralarm (code Rood)

De resultaten voor ‘Corners van SGG’ (Fig. A.1.3) zijn analoog aan die voor “Midden van SGG’.



Symbool	Uitleg
◆	Individuele provincies
▲	Alle provincies samen
■	Nederland

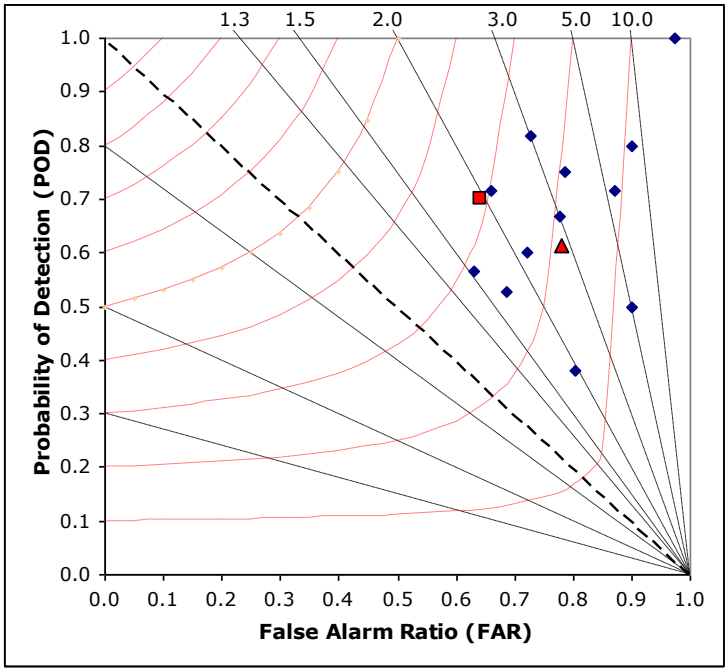
ALL		>300 ontladingen per 5 minuten in 50x50km vak	
		YES	**NO**
Weeralarm (Rood)	**YES**	25	58
	NO	151	15138

NL		>300 ontladingen per 5 minuten in 50x50km vak	
		YES	**NO**
Weeralarm (Rood)	**YES**	5	10
	NO	35	1928

	POD	FAR	F	CSI	Bias	A	B	C	D	N	SEDS	HKS	EDI	SEDI
ALL	0.14	0.70	0.00	0.11	0.47	25	58	151	15138	15372	0.509	0.138	0.481	0.491
NL	0.13	0.67	0.01	0.10	0.38	5	10	35	1928	1978	0.469	0.120	0.434	0.443

Fig. A1.3. Als in Fig. A1.1 maar dan voor ‘Corners van SGG’.

b. Waargenomen onweer (>300/5 min) vs waarschuwing voor extreem weer (code Oranje)



ALL		>300 ontladingen per 5 minuten in 50x50km vak	
		YES	**NO**
Waarschuwing (Oranje)	**YES**	108	384
	NO	68	14812

NL		>300 ontladingen per 5 minuten in 50x50km vak	
		YES	**NO**
Waarschuwing (Oranje)	**YES**	28	50
	NO	12	1888

	POD	**FAR**	**F**	**CSI**	**Bias**	**A**	**B**	**C**	**D**	**N**	**SEDS**	**HKS**	**EDI**	**SEDI**
ALL	0.61	0.78	0.03	0.19	2.80	108	384	68	14812	15372	0.593	0.588	0.766	0.800
NL	0.70	0.64	0.03	0.31	1.95	28	50	12	1888	1978	0.672	0.674	0.822	0.854

Fig. A1.4. Als in Fig. A1.1 maar dan voor code Oranje en ‘Corners van SGG’.

APPENDIX A2. Verificatie als functie van de forecasttijd voor 'Corners van SGG'

De tabellen en figuren voor NL zijn voor 'Corners van SGG' uiteraard hetzelfde als die voor 'Midden van SGG' en zijn hier herhaald om vergelijking met ALL makkelijker te maken. De conclusies bij de overeenkomstige plaatjes en tabellen zijn grotendeels gelijk aan die in hoofdstuk VI.2.

Code Rood

Corners ALL Rood 500	A	B	C	D	N	FAR	POD
0 hr	11	13	69	14701	14794	0.54	0.14
1 hr	13	12	67	17702	17794	0.48	0.16
2 hr	6	19	74	17696	17795	0.76	0.08
3 hr	0	24	80	14697	14801	1.00	0.00
4 hr	0	17	80	14714	14811	1.00	0.00
5 hr	0	8	80	14739	14827	1.00	0.00
6 hr	0	5	80	14749	14834	1.00	0.00

Corners NL Rood 500	A	B	C	D	N	FAR	POD
0 hr	2	3	16	1859	1880	0.60	0.11
1 hr	3	3	15	1859	1880	0.50	0.17
2 hr	1	5	17	1857	1880	0.83	0.06
3 hr	0	5	18	1859	1882	1.00	0.00
4 hr	0	3	18	1863	1884	1.00	0.00
5 hr	0	2	18	1868	1888	1.00	0.00
6 hr	0	1	18	1870	1889	1.00	0.00

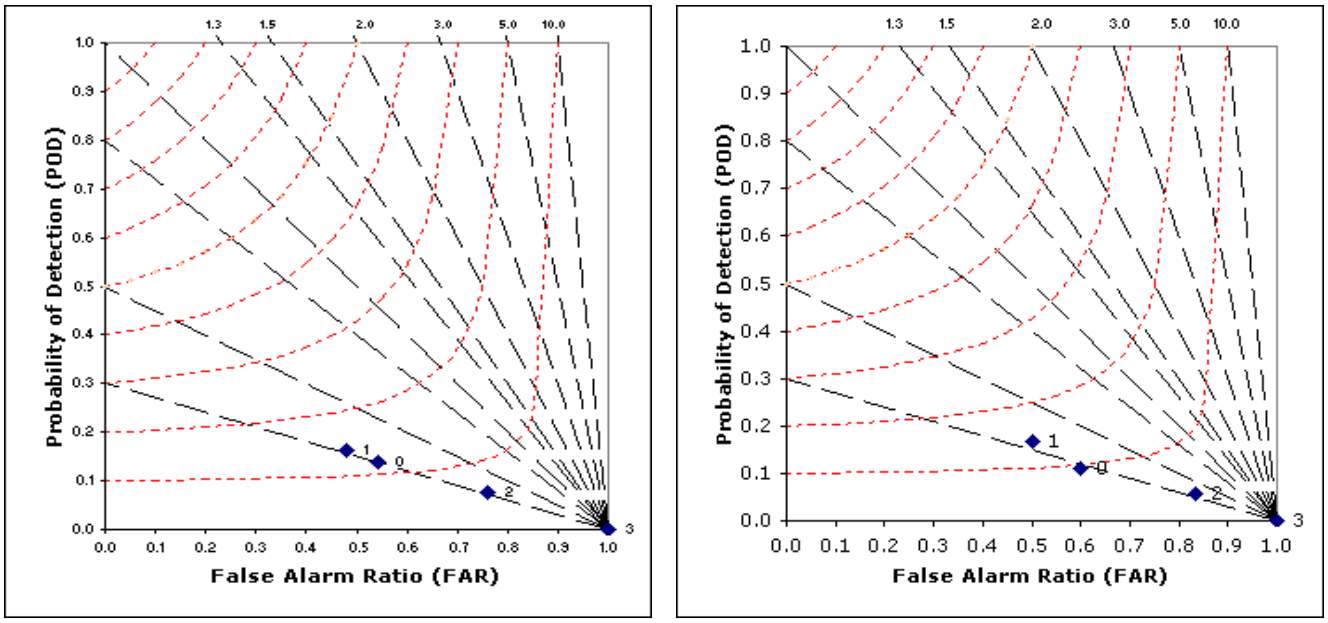
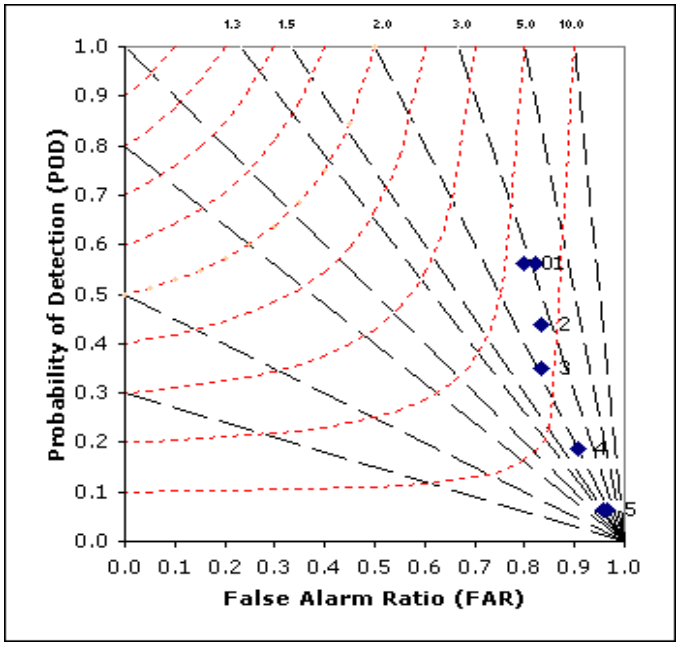


Fig. A2.1. POD-FAR plot voor code Rood waarschuwingen als functie van de forecasttijden vanaf +6 uur tot +0, met boven de plots de bijbehorende elementen van de contingentietabellen voor alle provincies samen (ALL, links) en Nederland (NL, rechts).

Code Oranje

Corners ALL Oranje 500	A	B	C	D	N	FAR	POD
0 hr	45	177	35	14537	14794	0.80	0.56
1 hr	45	210	35	14504	14794	0.82	0.56
2 hr	35	175	45	14540	14795	0.83	0.44
3 hr	28	140	52	14581	14801	0.83	0.35
4 hr	15	143	65	14588	14811	0.91	0.19
5 hr	5	136	75	14611	14827	0.96	0.06
6 hr	5	112	75	14642	14834	0.96	0.06

Corners NL Oranje 500	A	B	C	D	N	FAR	POD
0 hr	15	33	3	1829	1880	0.69	0.83
1 hr	14	40	4	1822	1880	0.74	0.78
2 hr	10	34	8	1828	1880	0.77	0.56
3 hr	7	25	11	1839	1882	0.78	0.39
4 hr	5	22	13	1844	1884	0.81	0.28
5 hr	3	19	15	1888	1925	0.86	0.17
6 hr	3	15	15	1889	1922	0.83	0.17



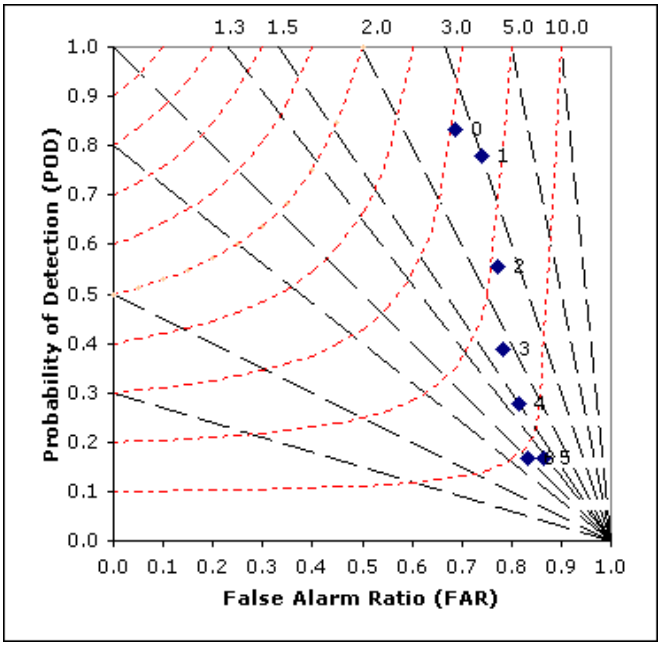


Fig. A2.2. Als in Fig. A2.1 maar dan voor code Oranje.