KNMI - Sector WM

MEMORANDUM

VOLGNUMMER	:	WM 00-10
ONDERWERP	:	Samenhang tussen kansen over een periode en die over deelperioden
SAMENSTELLER	:	Kees Kok
DATUM	:	november 2000

Dit is een persoonlijk memorandum. Slechts de auteur is verantwoordelijk voor de inhoud.
Indien niet meer nodig, graag retour auteur.

Verspreidingslijst

Bibliotheek
P&A
WM-AM
Meteorologen AMD
Wachtchefs AMD
Meteorologen LMD
Meteorologen Hoek van Holland
HAMD
C. Smith
E. Sluiter
B. Anker
C. Mei
D. Hart
Projectenbureau AMD

Samenhang tussen kansen over een periode en die over deelperioden

Kees Kok

Inhoudsopgave

1.	Inleiding	2
2.	Probleemstelling	2
3.	Enkele definities en notaties	4
4.	Berekening voor de verschillende verbanden tussen de deelgebeurtenissen	6
5.	Relatie met de covariantie tussen de deelgebeurtenissen	10
6.	Klimatologische illustratie	12
7.	Slotopmerkingen	15
Referenties		16

1. Inleiding

In de workshop Kansverwachtingen t.b.v. WA-meteorologen van de afgelopen zomer is een aantal sommetjes behandeld over de samenhang die er bestaat tussen kansen op bijvoorbeeld neerslag in een 24-uursperiode en de neerslagkansen die gegeven kunnen worden op deelintervallen. In het onderhavige stukje volgt een nadere uitwerking van dit probleem en wordt een aantal aspecten besproken waar rekening mee gehouden moet worden. De theorie zou oorspronkelijk in de vorm van een VOM-cursus behandeld worden en dus beschikbaar komen in een syllabus. Daar de cursus geen doorgang kon vinden is gekozen voor een vastlegging in de vorm van dit memo.

Allereerst wordt in dit pamflet nader ingegaan op de probleemstelling en de mogelijke toepasbaarheid van de zaken die hier vermeld zullen worden. Omdat in de workshop gebleken is dat de basiskennis omtrent kansen (definitie, etc.) grotendeels ontbrak en / of niet behandeld werd komt ook dit hier kort aan de orde (in paragraaf 3). De onderlinge samenhang tussen kansen over deelintervallen en de kans over de totale periode wordt geschetst allereerst in de vorm van de wiskundige restricties. De theorie wordt in paragraaf 4 uitgelegd aan de hand van een drietal voorbeelden die representatief zijn voor 3 min of meer standaardsituaties die als referentie kunnen dienen voor de meteoroloog bij het zelf bepalen van de onderlinge samenhang van de kansen. Hierbij wordt steeds uitgegaan van 2 disjuncte deelintervallen die tesamen de totale beschouwde periode beslaan. Een enkele keer wordt in het kort gekeken naar meer subintervallen. In sectie 5 wordt een uitstapje gemaakt naar wat wiskunde achter de zaak. Dit is niet direct noodzakelijk voor het begrijpen van de rest van het verhaal. In paragraaf 6 wordt een klimatologische illustratie gegeven van hoe de kansen, of liever de waargenomen frequenties, op een 11-jarige dataset er voor een drietal stations voor neerslag uitzien. We besluiten in paragraaf 7 met wat slotopmerkingen. Het idee van dit verhaal is deels ontleend aan een artikel van Krzysztofowicz uit 1999. Veel meer details zijn dan ook te vinden in dat artikel. Een andere belangrijke motivatie komt van de praktijk van de DeBav-kansverwachtingen die in het verleden niet altijd voldeden aan de in dit stuk beschreven randvoorwaarden.

2. Probleemstelling

Kansverwachtingen zullen in toenemende mate gebruikt (moeten) worden in de operationele praktijk. Temeer als we naar de situatie toe willen (en dat willen we) dat we ook uitspraken gaan doen over de betrouwbaarheid die we zelf toekennen aan individuele verwachtingen. In feite kunnen bijna alle verwachtingen opgevat worden als kans-verwachtingen (of zijn impliciet als zodanig bedoeld).

Voor het maken van kansverwachtingen beschikt de meteoroloog vaak niet over al te veel eenduidige kansinformatie uit de modellen. Deterministische modellen geven geen kansen. Voor sommige predictands (de te voorspellen meteorologische grootheden) waarvoor kansen gegeven moeten worden geven deze modellen wel informatie die gebruikt kan worden, maar alleen in de vorm van gridboxgemiddelden. Dit is bijvoorbeeld het geval voor neerslag en bewolking. De meteoroloog moet dit vaak zelf vertalen naar een kans dat een willekeurige persoon, of een willekeurig gebied, etc. met het fenomeen in aanraking komt. Deterministische modellen leveren geen rechtstreekse informatie over bijv. de kans op mist, onweer of hagel. De verwachte synoptische situatie en de verticale opbouw van de modelatmosfeer helpen de meteoroloog de kans op deze verschijnselen te bepalen. Hoewel er

hierbij gebruik gemaakt wordt van modeluitvoer wordt het op deze manier verkrijgen van kansen "subjectief" genoemd.

Voor de middellange termijn is de situatie voor een beperkt aantal predictands iets gunstiger. EPS, op zichzelf een verzameling deterministische modelruns, levert door middel van zijn 50 oplossingen die min of meer even waarschijnlijk verondersteld worden, een schatting van iets dat als kans opgevat kan worden. Maar ook hier weer voor een zeer beperkte set predictands. Daarnaast is deze methode niet bruikbaar voor de termijn korter dan 3 dagen vooruit.

Er zijn voor een aantal elementen echter ook "objectieve" kansen beschikbaar in de vorm van zogenaamde gidsen. Hierbij zijn op grond van een historische dataset van modelverwachtingen en bijbehorende waarnemingen "objectieve" kansen berekend voor de verschijnselen. Voor m.n. de kans op neerslag, onweer en windsterkte boven 6 en 8Bf wordt hiervan op het KNMI al zo'n 20 jaar gebruik gemaakt. Een niet te onderschatten voordeel van gidskansen is dat deze zodanig afgeleid zijn dat ze in principe *reliable* zijn. Dit wil zeggen dat bij kansuitspraken van bijv. p% op het optreden van een bepaald verschijnsel in principe ook in p% van de gevallen het verschijnsel inderdaad zal optreden. De set van objectieve kansverwachtingen is echter lang niet compleet: niet voor elk verschijnsel op elke plaats, tijdsinterval, forecastperiode is er een gids beschikbaar. De forecaster moet zelf vaak kansen inschatten of de beschikbare "gidskansen" aanpassen.

Dit verhaal gaat niet over het (subjectief) bepalen van de kansverwachtingen uit modellen, noch over wanneer en hoe gidskansen aangepast moeten worden. Maar het gaat wèl over welke randvoorwaarden er zijn, wiskundig gezien en meteorologisch gezien, waar rekening mee gehouden moet worden bij het schatten of aanpassen van kansen die met elkaar zouden kunnen samenhangen. Dit is, meer in het bijzonder, ook van toepassing bij het combineren van objectieve en subjectieve kansen. Alleen de samenhang tussen kansen over een bepaalde periode en de kansen over disjuncte deelintervallen die de totale periode beslaan worden bekeken. Ter focussering van de gedachten wordt alleen gesproken over de kans op neerslag. Maar de conclusies zijn uiteraard onafhankelijk van de gekozen predictand.

Als voorbeeld kan gedacht worden aan de situatie dat de MOS (Model Output Statistics) gidsverwachting een kans op neerslag in een 24-uursperiode geeft, en dat de meteoroloog moet bepalen wat de kans op neerslag 's ochtends en 's middags (apart) is. Dit kan hij / zij bijvoorbeeld doen aan de hand van additionele informatie over het geschatte tijdstip van overkomen van een front of over de mogelijkheid van het ontstaan van buien. Ander voorbeeld: er komt een gids voor een overschrijdingskans op (grote) hoeveelheden neerslag ergens in de verwachtingstermijn van dag 6 tot dag 10. Wat is dan een realistische kans voor de individuele dagen (met of zonder extra synoptische informatie of extra informatie over de correlatie tussen de kansen op de verschillende dagen)? Een laatste voorbeeld, afkomstig uit een niet al te ver verleden, is het volgende. Voor DeBav moesten indertijd neerslagkansen gegeven worden voor een aantal achtereenvolgende perioden van 3 uur. De enige informatie die beschikbaar was was een op MOS gebaseerde 24-uurskans en de neerslagprognoses van het LAM en de daarbij horende verwachte synoptische situatie. Hoe moeten deze kansen op een onderling consistente manier bepaald worden zodanig dat ook nog recht gedaan wordt aan de skill van de (uiteindelijke) verwachting? Dit soort vragen was in het kort (?) de achterliggende reden van de sommetjes van de workshop. In de volgende paragrafen wordt een aantal argumenten cq. achtergronden gegeven die kunnen helpen bij het beantwoorden van deze vragen. Maar eerst volgen nog enkele definities en notaties die gebruikt worden.

3. Enkele definities en notaties

Hoewel het intuitieve begrip van kansen in het algemeen meer dan voldoende is, is het soms toch handig om een exacte definitie te hanteren. Er zijn veel kansdefinities in omloop, die, hoewel niet toepasbaar op alle problemen, wel allemaal leiden tot dezelfde rekenregels. Een veel gebruikte, die in ieder geval het onthouden van de rekenregels makkelijk maakt, is de kansdefinitie volgens Lapace. Deze definieert de kans op een gebeurtenis als de verhouding tussen het aantal gunstige gevallen en het totaal aantal gevallen. Deze verhouding kan op theoretische gronden of door middel van experimenten (zoals bijv. bij het 1000 maal gooien met een onzuivere dobbelsteen) bepaald of geschat worden. In veel gevallen en zeker in de meteorologie kunnen we dit niet zo simpel vertalen; we kunnen bijvoorbeeld geen experimenten doen om de kans op een gebeurtenis te bepalen. Hier wordt dan ook vaak de onzekerheid over de deterministische modelverwachting als een van de criteria genomen voor het bepalen van de kans. Voor meer details over de kansdefinities zie bijv. Buijs (1997).

In dit pamflet beschouwen we kansen op neerslag. De kansen worden genoteerd met P (afkomstig van probability), de beschouwde gebeurtenis (neerslag) met E (event). Maar neerslag wordt hier gedefinieerd als een hoeveelheid groter dan 0 mm en niet als tenminste 0.3 mm (zie later). De kansen op neerslag in de disjuncte deelintervallen worden genoteerd als p_1 en p_2 . Deze subevents, i.e. neerslag in deelinterval 1 en 2, worden E_1 en E_2 genoemd. Voor het gemak nemen we voor de totale periode 24 uur (00 tot 24GMT) en voor de deelperioden de eerste en de tweede 12 uursperiode. Deze worden in het vervolg in het kort ook ochtend en middag genoemd. Dus:

E = een gemeten neerslaghoeveelheid van > 0 mm in de 24-uursperiode
E_1 idem in de eerste 12 uur van deze periode,
E_2 in de tweede.
$p = P(E)$, is de kans op E,
$p_1 = P(E_1)$, de kans op E_1;
analoog voor p_2.

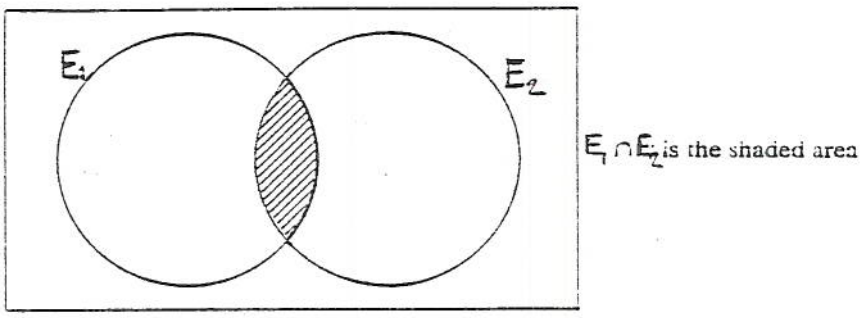
Er geldt dat alle kansen liggen tussen 0 en 1 (incl. 0 en 1). Verder:

$p \geq p_1$ en $p \geq p_2$, Liever nog:
$p \geq \max(p_1, p_2)$
$p \leq p_1 + p_2$ of nog liever
$p \leq \min(p_1+p_2, 1)$

Deze regels zijn meteen evident als de kansdefinitie van Laplace gehanteerd wordt. Maar, zoals gezegd, gelden ze voor alle kansdefinities.
Tevens kunnen kansen gegeven worden voor het optreden van beide events (zowel 's ochtends als 's middags neerslag) en voor tenminste één van beide events, resp. $P(E_1 \cap E_2)$ en $P(E_1 \cup E_2)$. Deze laatste is in onze notatie gelijk aan p.

N.B. Let wel dat we hier neerslag genomen hebben als er meer dan niks gevallen is. Als we zoals gangbaar is neerslag definiëren als er tenminste 0.3 mm gevallen is, dan hoeft de relatie $p \leq \min(p_1+p_2, 1)$ niet meer te gelden. Er kan zich dan namelijk het verschijnsel voordoen dat het zowel 's ochtends als 's middags 0.2 mm geregend heeft en dus "droog" was, terwijl er op die dag tesamen dus meer dan 0.3 mm afgetapt werd. Dit is de reden voor de boven gehanteerde definitie. Dit probleem doet zich niet voor bij de meeste andere predictands waarvoor kansverwachtingen gemaakt moeten worden (onweer, mist, etc.). Maar ook bij neerslag is dit in verreweg de meeste gevallen te verwaarlozen.

Voor het rekenen met twee events is het gebruik van het bekende Venn-diagram illustratief (voor meer events loopt dit al gauw spaak). Zie de schets hieronder. Hierin zijn binnen de totale oplossingsruimte (het geheel binnen de rechthoek) de verzamelingen van het optreden van de 2 gebeurtenissen schematisch weergegeven. De totale oplossingsruimte is het totaal aantal gevallen met al of niet neerslag op de hele dag. Ovaal E_1 betreft de gevallen met 's ochtends neerslag, ovaal E_2 die met 's middags. De doorsnede in ons probleem betreft dus de gevallen waarin in beide deelintervallen neerslag gevallen is (> 0 mm). Let op dat de gebeurtenissen disjunct zijn in de tijd maar dat ze in het Venn-diagram een niet-lege doorsnee kunnen hebben.



Een belangrijke regel die onmiddellijk duidelijk is uit het diagram is:

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) \qquad (1a)$$

In onze korte notatie is dit:

$$p = p_1 + p_2 - P(E_1 \cap E_2). \qquad (1b)$$

Dit is de algemeen geldende relatie tussen de kans over het hele interval en de kansen over zijn disjuncte deelintervallen.
De waarde van $P(E_1 \cap E_2)$ bepaalt dus in belangrijke mate de onderlinge verhouding tussen de kans over het hele interval en die over deelintervallen. In de volgende paragraaf beschouwen we drie zeer uiteenlopende meteorologische omstandigheden die aanleiding geven tot drie zeer verschillende waarden voor deze doorsnede.

Tot slot nog enkele opmerkingen over **voorwaardelijke** (of conditionele) **kansen**. In het bovenstaande werd steeds de hele uitkomstenruimte bekeken. Vaak is er voor de bepaling van de kans veel meer informatie. Vergelijk: "wat is de kans om met een zuivere dobbelsteen een 6 te gooien" met "wat is de kans op een 6 gegeven dat het een even getal is". Deze extra informatie verdubbelt hier dus de kans. De (voorwaardelijke) "kans op E_2 gegeven E_1" , genoteerd als $P(E_2 \mid E_1)$, is hier dus de kans op neerslag in de tweede 12-uursperiode gegeven dat er neerslag gevallen is in de eerste 12-uursperiode. Deze grootheid kan dus gezien worden als een maat voor het persisterende (of repeterende) karakter van de neerslag. We zullen dan ook zien in de volgende paragraaf dat de grootte van deze term sterk samenhangt met de synoptische situatie. Uit het Venn-diagram is duidelijk dat bij $P(E_2 \mid E_1)$ alleen gekeken wordt naar de gevallen met 's ochtends neerslag en van die gevallen alleen het percentage berekend wordt waarbij er 's middags neerslag gevallen is. Er geldt dus

$$P(E_2 \mid E_1) = P(E_1 \cap E_2) / P(E_1) \qquad (2)$$

Hierbij nemen we voor het gemak aan dat $P(E_1) = p_1$ ongelijk aan 0 is. We zullen in de situaties die in paragraaf 4 besproken worden ook de waarden van de voorwaardelijke kansen betrekken.

4. Berekening voor verschillende verbanden tussen de deelgebeurtenissen

In deze paragraaf beschouwen we 3 verschillende afhankelijkheden tussen de ochtend- en middagevent en de consequenties hiervan voor de dagevent.

Situatie a). *de gebeurtenissen sluiten elkaar uit*

Het vóórkomen van de een resulteert in het niet vóórkomen van de ander. I.e.

$P(E_1 \mid E_2) = 0$ en $P(E_2 \mid E_1) = 0$

In het Venn-diagram zijn de twee verzamelingen disjunct: er zijn geen gevallen waarin het zowel 's ochtends als 's middags regent. Dus $P(E_1 \cap E_2) = 0$.
Deze situatie doet zich bijvoorbeeld voor wanneer er een smal front verwacht wordt, waarbij het niet zeker is of het 's ochtends of 's middags passeert en waarbij we voor het gemak even aannemen dat het rond het middaguur in ieder geval droog is. Het neerslaggebied moet bijvoorbeeld dusdanig smal zijn dat het rond het middaguur de positie waarvoor de kansverwachting gemaakt moet worden al gepasseerd is of nog moet bereiken. Een ander geval dat deze situatie benadert is het geval waarin er een kans is op een bui en waarbij àls er een bui valt deze het convectiemechanisme teniet doet en de ontwikkeling van volgende buien onmogelijk maakt. Is er 's ochtends geen bui gevallen dan blijft er een kans dat het 's middags gebeurt. In deze voorbeelden geldt dan (zie (1)):

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

ofwel

$$p = p_1 + p_2 \ (\leq 1).$$

Doet zo'n situatie van grote negatieve afhankelijkheid zich voor dan moet opgelet worden dat p_1 en p_2 samen niet groter zijn dan 1. Een veel grotere kans dat de veronderstelde bui of frontpassage 's ochtends voorkomt impliceert dan een zeer kleine kans voor de middag.

Situatie b). *de gebeurtenissen impliceren elkaar*

Het vóórkomen van de ene gebeurtenis impliceert het vóórkomen van de andere gebeurtenis. Ofwel:

$P(E_1 \mid E_2) = 1$ en $P(E_2 \mid E_1) = 1$.

De kans op de een is dan niet alleen gelijk aan de kans op de ander, maar ook de kans op de vereniging en de kans op de doorsnee zijn hieraan gelijk. In het Venn-diagram overlappen E_1 en E_2 elkaar volledig. I.e.

$p = p_1 = p_2$

Dit geval doet zich bijvoorbeeld voor in de situatie dat er een al of niet grote kans is dat er een breed front in de morgen het land binnendringt en dan gedurende een groot deel van de rest van de dag regen zal brengen, maar dat er ook een kans bestaat dat de koers niet richting Nederland is en het front ons land helemaal niet zal bereiken. Het voorspelprobleem (wat betreft kansen) is in dit geval teruggebracht tot het bepalen van slechts één kans. Als er al een objectief bepaalde neerslagkans beschikbaar is, bijvoorbeeld verkregen via MOS, èn de meteoroloog voelt geen behoefte deze te veranderen, dan is de zaak rond. Is er bijvoorbeeld een MOS-kans van 60% dat er 's ochtends neerslag komt, dan is de kans voor de middag en voor de hele dag hieraan gelijk.

Een uitbreiding op dit laatste voorbeeld, die strikt genomen niet tot deze categorie behoort, wordt gegeven na de bespreking van situatie c).

Situatie c). *de gebeurtenissen zijn onafhankelijk van elkaar*

De gebeurtenissen kunnen dus los van elkaar beschouwd worden. Dit komt voor bij bijvoorbeeld buiig weer waarbij er meerdere buien per dag kunnen vallen zonder dat het vallen van een bui de kans beïnvloedt op nog een bui op dezelfde plaats. Of als het weer 's ochtends bepaald wordt door een warmtefront terwijl 's middags het weer een buiig karakter krijgt. De kans dat er tenminste 0.1 mm neerslag valt op een dag, of dat nu 's ochtends is of 's middags of beide, wordt dan bepaald "door het toeval". De kans dat op een dagdeel neerslag valt is dan niet meer afhankelijk van of er op het andere dagdeel juist niet of juist wel neerslag gevallen is. Dus voor de conditionele kansen geldt

$P(E_2 \mid E_1) = P(E_2)$ en $P(E_1 \mid E_2) = P(E_1)$.

M.a.w. voor de kans op de ene gebeurtenis doet de extra informatie over het al of niet vóórkomen van de andere er niets toe. Dit in tegenstelling tot de gevallen a) en b). Uit formule (2) volgt dan de bekende formule voor onafhankelijke gebeurtenissen:

$P(E_1 \cap E_2) = P(E_1) \,.\, P(E_2)$

Hiermee wordt (1):

$$p = p_1 + p_2 - p_1.p_2$$

Generalisatie naar meer dan 2 deelintervallen leidt tot een heel wat minder aantrekkelijke uitdrukking. Daarom geven we ook een alternatieve aanpak. Bovenstaande kan namelijk ook "afgeleid" worden door te kijken naar de complementen, i.e. we beschouwen niet de kans op neerslag maar juist de kans dat het de hele dag droog zal blijven. Dit impliceert dat beide deelintervallen droog moeten zijn. De kans op een droge dag is dan $(1 - p_1).(1 - p_2)$; de kans op een niet-droge dag is dus $1 - (1 - p_1).(1 - p_2)$. Dit is inderdaad gelijk aan $p_1 + p_2 - p_1.p_2$.

Voor meer deelintervallen, zeg k, geldt

$$p = 1 - (1 - p_1).(1 - p_2) \ldots (1 - p_k) \,. \qquad (3)$$

Dit is dus het verband tussen p en p_i voor het geval dat alle deelevents onafhankelijk van elkaar zijn.

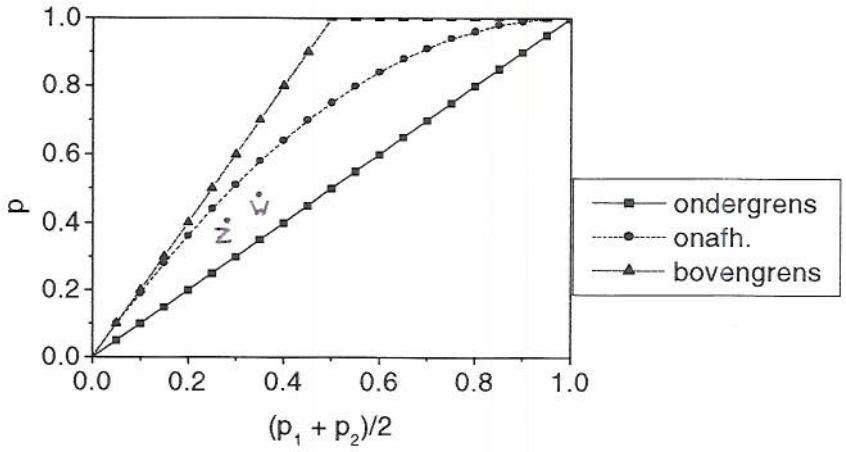
□

De bovenstaande 3 situaties, a) tot en met c), zijn min of meer geïdealiseerde uitersten. In de praktijk zullen we vaak situaties tegenkomen die met geen van de drie exact in overeenstemming zijn maar misschien wel in meer of mindere mate lijken op één (of een combinatie) van deze 3, of ongeveer tussen 2 van de beschreven situaties inliggen.
Als voorbeeld geven we een situatie die lijkt op het bovenbeschreven geval b), maar nu zodanig dat nog maar één gebeurtenis de ander impliceert en niet meer ook de ander de één. (In de praktijk betekent dat natuurlijk dat de gebeurtenis E_2 een gevolg is van E_1 en niet andersom). We bekijken hiervoor net als in b) weer een grootschalig regengebied, bijv. horend bij een warmtefront, dat àls het regen brengt langdurig regen brengt. De eventuele regen die 's ochtends begint houdt dan aan tot in de middag. Maar stel dat nu het front niet alleen de opties heeft (zoals in b)) om òf Nederland binnen te trekken in de ochtend òf om Nederland helemaal niet aan te doen, maar dat het nog een derde optie heeft, namelijk om àls Nederland aangedaan wordt een ietsje later te arriveren zodat het 's ochtends nog droog blijft maar 's middags alsnog regen geeft. In dat geval zal p_1 nu kleiner zijn dan p_2 , maar deze laatste is even groot als p. Wordt bijvoorbeeld de kans op het bereiken van het front hoog geacht, zeg 80%, met een grote waarschijnlijkheid dat het al 's ochtends arriveert, dan zouden de waarden voor p_1, p_2 en p bijvoorbeeld 60%, 80% en 80% kunnen zijn. In het algemeen geldt voor de hier beschreven situatie:

$$p_1 \leq p_2 = p$$

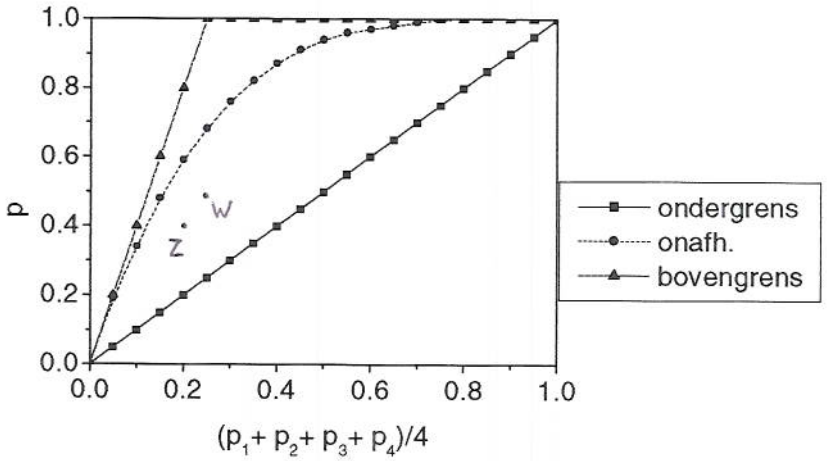
Het verband tussen de kans over het totale interval en die over de subintervallen kan voor de bovenstaande 3 "modellen" ook grafisch weergegeven worden; zie onderstaande grafiek. Om een 2-dimensionale presentatie mogelijk te maken is hierin p uitgezet tegen de gemiddelde waarde van p_1 en p_2 (i.p.v. aparte assen voor p_1 en p_2). De waarden van p liggen dus bij gegeven kansen op de deelintervallen p_i , in het relatief smalle gebied tussen extreme positieve afhankelijkheid tussen de gebeurtenissen 's ochtends en 's middags (i.e. de diagonaal) en extreme negatieve afhankelijkheid ($p = p_1 + p_2$). Deze laatste heeft uiteraard een maximum van 1. Deze 2 lijnen komen dus overeen met resp. situatie b) en situatie a). Zijn de gebeurtenissen onafhankelijk dan volgt p de kromme lijn. Voor deze kromme èn voor de lijn die de ondergrens aangeeft is in dit plaatje nog een extra aanname gemaakt en wel dat p_1 en p_2 hier gelijk genomen zijn. Zijn deze 2 kansen ongelijk (maar met de som nog hetzelfde) dan liggen deze krommen iets hoger en is de marge van mogelijke waarden van p nog kleiner. Zo wordt de ondergrens uiteraard bepaald door de voorwaarde dat $p \geq \max(p_1, p_2)$ en dit is bij ongelijke p_i's groter dan $(p_1+p_2)/2$. Voor het onafhankelijke geval volgt het rechtstreeks uit de formule gegeven bij geval c). De bovengrens, bepaald door de voorwaarde dat $p \leq \min(1,p_1+p_2)$, blijft ongewijzigd bij ongelijke p_i's. Voor kleine waarden van p_i liggen de kansen voor het totale interval vlakbij de maximale waarde.

Het gebied boven de gebogen kromme (i.e. voor onafhankelijke deelevents) wordt gevormd door de gevallen waarin er een negatieve correlatie is tussen de twee events, er onder door positief gecorreleerde deelevents. Zie ook paragraaf 5. Een figuur als dit stelt de forecaster in staat afhankelijk van de geschatte afhankelijkheid tussen E_1 en E_2 de onderlinge samenhang tussen de 3 kansen te bepalen. De punten aangegeven in de figuur door 'Z' en 'W' worden besproken in paragraaf 6.



Bovenstaande kan gemakkelijk gegeneraliseerd worden tot meer deelintervallen. Nemen we bijvoorbeeld 4 deelintervallen dan krijgen we een figuur zoals hieronder. Wederom is p hier uitgezet tegen het gemiddelde van de kansen op de deelevents. De ondergrens wordt weer bepaald door geval b), de bovengrens door a). De kromme voor de onafhankelijke situatie en die voor de ondergrens zijn weer getekend voor het geval dat alle p_i's gelijk zijn. Zijn ze niet allemaal gelijk dan liggen deze lijnen weer iets hoger. De punten aangegeven in de figuur door 'Z' en 'W' worden besproken in paragraaf 6.

Als getallenvoorbeeld beschouwen we de situatie met een buiig weertype waarbij de buiigheid over de beschouwde periode niet toe- of afneemt. Dus voor de locatie waar de verwachting voor moet gelden, is op ieder moment in de periode de kans op neerslag hetzelfde. Dit is dus het boven beschreven geval c). Stel we hebben een MOS-kans op neerslag voor de 24-uursperiode of een inschatting van de meteoroloog. Dus we hebben een gegeven PoP24, i.e. de kans op neerslag (Probability of Precipitation) in een 24-uursperiode, met hier weer neerslag gedefinieerd als 0.1 mm of meer. Stel bijvoorbeeld PoP24 = 75%, een vrij hoge neerslagkans. Dan volgt voor resp. 2 en 4 disjuncte deelintervallen dat PoP12 = 50%, PoP6 ~ 30% (zie de 2 figuren). Een uurlijkse neerslagkans zou op analoge manier onder de bovenstaande aannames uitkomen op bijna 6%.



5. Relatie met de covariantie tussen de deelgebeurtenissen

In deze paragraaf maken we een wiskundig uitstapje en bekijken we nogmaals de formulering van de relatie tussen de kans op de gebeurtenis in het totale interval met de kansen op de gebeurtenis in zijn deelintervallen. Dus formule (1), maar in de vorm van onze verkorte notatie:

$p = p_1 + p_2 - P(E_1 \cap E_2)$.

We beschouwen weer voor het gemak de gebeurtenis van het al of niet optreden van neerslag en we bekijken alleen de situatie met 2 disjuncte deelintervallen die tesamen het totale interval beslaan.

Om het verband beter te kunnen kwantificeren kennen we nu eerst getallen toe aan het wel en niet vóórkomen van de events. We definiëren een $x1$ voor de gebeurtenis E_1 , waarbij $x1 = 1$ als er tenminste 0.1 mm neerslag valt in het eerste deelinterval, zeg 's ochtends, en $x1 = 0$ als het droog blijft. Analoog $x2$ voor de tweede deelperiode, 's middags, en x voor de hele dag. Hiermee is het gemiddelde aantal gevallen met neerslag in de 24-uursperiode gelijk aan

$\frac{1}{n}\sum_{i=1}^{n} x_i$, met n het totaal aantal gevallen.

Dit gemiddelde wordt ook wel weergegeven door $\bar{x}$. Analoog kunnen gemiddelden gegeven worden voor $x1$ en $x2$. Deze gemiddelden (ofwel frequenties) kunnen, als het aantal gevallen dat beschouwd wordt groot genoeg is, geïnterpreteerd worden als kansen op de gebeurtenissen. Met andere woorden, $\bar{x} = p$, $\overline{x1} = p_1$ en $\overline{x2} = p_2$. Er volgt dan voor het verband tussen het vóórkomen van neerslag over de hele dag met het vóórkomen van neerslag op de ochtend en dat op de middag de volgende relatie:

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} x1_i + \frac{1}{n}\sum_{i=1}^{n} x2_i - \frac{1}{n}\sum_{i=1}^{n} (x1.x2)_i$$

Deze formule is het equivalent van formule (1) en volgt dan ook rechtstreeks uit het Venn-diagram aldaar. In onze oorspronkelijke notatie:

$$p = p_1 + p_2 - \frac{1}{n}\sum_{i=1}^{n} (x1.x2)_i$$

met de laatste term dus gelijk aan $P(E_1 \cap E_2)$.

We gaan nu gebruik maken van de definitie van de **covariantie**. Dit is een maat voor de samenhang tussen twee grootheden en is gedefinieerd als volgt:

$$\mathrm{cov}(x1,x2) = \frac{1}{n}\sum_{i=1}^{n} (x1_i - \overline{x1}).(x2_i - \overline{x2}) \qquad (4a)$$

ofwel

$$\mathrm{cov}(x1,x2) = \frac{1}{n}\sum_{i=1}^{n} (x1.x2)_i - \overline{x1}\,\overline{x2} \qquad (4b)$$

Uit de bovenste formulering kunnen we zien dat de covariantie een kwantificering geeft van de mate waarin, gemiddeld genomen, de grootheden $x1$ en $x2$ tegelijkertijd boven het gemiddelde dan wel onder het gemiddelde optreden. Bij positieve covariantie fluctueren ze dus in meer of mindere mate in fase. Bij ongecorreleerde (onafhankelijke) grootheden is de covariantie gelijk aan 0. Het begrip covariantie is nauw gerelateerd aan de correlatie. Deze laatste is een normalisering van de eerste en heeft dus hetzelfde teken.

In de notaties van de vorige paragraaf ziet formule (4b) er dan als volgt uit:

$$P(E_1 \cap E_2) = p_1.p_2 + \mathrm{cov}(x1,x2)$$

en kan het verband tussen p en de p_i's dus geschreven worden als

$$p = p_1 + p_2 - p_1.p_2 - \mathrm{cov}(x1,x2) \qquad (5)$$

□

We kunnen nu de covarianties uitrekenen voor de 3 gevallen uit paragraaf 4. Dit resulteert uiteraard in de oplossingen die aldaar gegeven zijn, maar nu met een explicicte uitdrukking voor de covariantie. Het is het gemakkelijkst om de formulering van de covariantie te gebruiken die vermeld staat in (4b).

In **situatie a).**, i.e. de ochtend- en middaggebeurtenissen sluiten elkaar uit, wordt de sommatie-term in (4b) gelijk aan 0 omdat combinaties van $x1$ en $x2$ alleen maar voorkomen in de vorm (0,1) en (1,0). Dit resulteert in

$$\mathrm{cov}(x1,x2) = -p_1.p_2$$

en dus in

$$p = p_1 + p_2.$$

Situatie b). uit paragraaf 4 betrof het geval waarbij de 2 gebeurtenissen elkaar impliceren. Er komen dan alleen combinaties (1,1) en (0,0) voor. De relatieve frequenties van voorkomen van de ochtend- en middaggebeurtenis zijn dan gelijk. De eerste term uit het rechterlid van (4b) wordt dan

$$\frac{1}{n}\sum_{i=1}^{n} x1_i.x2_i = \frac{1}{n}\sum_{i=1}^{n} (x1_i)^2 = \frac{1}{n}\sum_{i=1}^{n} x1_i = p_1$$

Dit resulteert in

$$\mathrm{cov}(x1,x2) = p_1 - p_1.p_2 = p_1 - (p_1)^2$$

en

$$p = p_1 = p_2 .$$

Situatie c). tenslotte, waarin de gebeurtenissen onafhankelijk zijn van elkaar, levert zoals we al gezien hebben een covariantie gelijk aan 0. Hiermee wordt dus

$$p = p_1 + p_2 - p_1.p_2 .$$

□

Tot slot nog een opmerking over **conditionele kansen**. Hierbij gaan we uit van formule (2) , i.e. $P(E_2 \mid E_1) = P(E_1 \cap E_2) / P(E_1)$, en gebruiken voor $P(E_1 \cap E_2)$ de expressie die we boven hebben afgeleid. Dit resulteert in

$$P(E_2 \mid E_1) = \{p_1.p_2 + \mathrm{cov}(x1,x2)\} / p_1$$

ofwel

$$P(E_2 \mid E_1) = p_2 + \mathrm{cov}(x1,x2) / p_1$$

Hieruit volgt dus dat als de correlatie (of covariantie) gelijk is aan 0, i.e. bij onafhankelijke gebeurtenissen

$P(E_2 \mid E_1) = P(E_2)$

Dit hebben al gezien in paragraaf 4.

Verder geldt dat

$P(E_2 \mid E_1) > P(E_2)$ als de correlatie tussen de deelevents positief is, en

$P(E_2 \mid E_1) < P(E_2)$ bij een negatieve correlatie.

Dit zijn relaties die van pas kunnen komen bij de interpretatie van de klimatologische kansen die in de volgende paragraaf behandeld worden.

6. Klimatologische illustratie

In deze paragraaf zullen we bekijken in hoeverre, gemiddeld gezien, het voorkomen van neerslag op delen van de dag met elkaar en met de neerslagkans over de hele dag te maken heeft. In dit geval hebben we het, in tegenstelling tot in de vorige paragrafen, over neerslag als de hoeveelheid minstens 0.3 mm bedraagt; minder dan 0.3 mm noemen we droog. We hebben de periode bekeken van 1 december 1989 tot 1 augustus 2000. In totaal dus ongeveer 3900 gevallen (dagen). De gemiddelde frequentie van de neerslag over de betreffende (delen van) dagen interpreteren we voor het gemak wederom als de kans. We hebben de data gesplitst in een zomer- en winterhalfjaar, resp. lopend van 1 april tot aan 1 oktober en van 1 oktober tot 1 april. We hebben een "kuststation" (De Kooy, 06235) vergeleken met twee "landstations" (De Bilt, 06260, en Beek, 06380).

Deze splitsing is gemaakt omdat neerslag naast een grootschalige component, die geen functie van het tijdstip op de dag is, en dus op een voldoende grote dataset geen verschillen te zien geeft, een belangrijke component heeft die wèl duidelijk een dagelijkse gang vertoont. Deze laatste is een functie van het seizoen en van de afstand tot de kust. Voor het (subjectief) inschatten door de meteoroloog van de kans op het voorkomen van een grootheid die soortgelijk gedrag zou kunnen vertonen is het van groot belang om de klimatologische frequentie te kennen. In het algemeen geldt uiteraard dat bij het maken van kansverwachtingen de meteoroloog altijd (bijvoorbeeld als referentie) de klimatologische kans als functie van de locatie en de tijd van het jaar dient te kennen.

Net als in paragraaf 4 onderscheiden we de opsplitsing in twee en vier subintervallen.

A. We kijken allereerst naar de opdeling van de dag **in 2 deelintervallen**. De kansen (percentages) over de hele dag alsmede over de ochtend (hier: 00-12) en de middag (hier: 12-24) zijn gegeven. Verder de voorwaardelijke kans op neerslag 's middags gegeven dat er 's ochtends neerslag gevallen is ($P(E_2 \mid E_1)$) en de kans dat het zowel 's ochtends als 's middags regende ($P(E_1 \cap E_2)$). Om een vergelijking met de formules uit de vorige paragrafen mogelijk te maken is ook de covariantie tussen (al of niet) neerslag 's ochtends en 's middags gegeven. Ook is de correlatie tussen neerslag 's ochtends en 's middags opgenomen, alsmede de voorwaardelijke kans op neerslag in de ochtend gegeven het feit dat er neerslag opgetreden is in de voorafgaande middag, genoteerd als $P(E_1 \mid E_2^{-1})$. Alle kansen en de waarde voor de covariantie zijn vermenigvuldigd met 100.

Omdat het aantal gevallen dat gebruikt is voor de 2 tabellen nagenoeg gelijk is kunnen jaarcijfers gemakkelijk berekend worden door de betreffende getallen uit de zomer- en wintertabel te middelen.

Zomerhalfjaar

	De Kooy	De Bilt	Beek
p1 (00-12)	27.7	26.9	24.0
p2 (12-24)	25.7	29.3	30.0
p (00-24)	40.1	40.6	39.5
$P(E_2 \mid E_1)$	48.9	58.7	60.9
$P(E_1 \cap E_2)$	13.5	15.8	14.6
covariantie	6.4	7.9	7.4
correlatie	0.329	0.393	0.378
$P(E_1 \mid E_2^{-1})$	56.5	53.5	45.1

Winterhalfjaar

	De Kooy	De Bilt	Beek
p1 (00-12)	34.1	33.6	32.0
p2 (12-24)	35.3	35.2	33.0
p (00-24)	49.8	48.1	46.5
$P(E_2 \mid E_1)$	59.3	62.8	58.8
$P(E_1 \cap E_2)$	20.2	21.0	18.8
covariantie	8.2	9.3	8.3
correlatie	0.360	0.410	0.377
$P(E_1 \mid E_2^{-1})$	59.8	58.0	56.5

Uit de tabellen volgt dat er in de laatste 11 jaar in De Kooy gemiddeld net iets vaker (niet noodzakelijkerwijs méér) neerslag gevallen is dan in de andere stations, en in Beek het minst vaak. Voor alle stations geldt dat in het winterhalfjaar neerslag vaker voorkomt, en ook vaker 's ochtends èn 's middags (i.e. $P(E_1 \cap E_2)$). In dat "jaargetijde" is er slechts een lichte voorkeur voor de middag boven de ochtend, (let op dat in onze data de 'winter' een half jaar lang is). In de "zomer" daarentegen is er voor de landstations wel een duidelijk hogere neerslagfrequentie 's middags dan 's ochtends; in Beek nog weer aanzienlijk hoger dan in De Bilt. In De Kooy daarentegen is het juist omgekeerd. Voor De Bilt is zowel voor de 'zomer' (Z) als de 'winter' (W) het verband tussen de kansen p_1 en p_2 met p weergegeven in de desbetreffende grafiek in paragraaf 4. Deze punten liggen in het gebied tussen het onafhankelijke geval en de ondergrens, duidend op een positieve correlatie tussen de deelevents (Ook als je in ogenschouw neemt dat deze twee lijnen voor de hier beschouwde p_i's ietsje hoger liggen dan in de figuur, a.g.v. het feit dat p_1 en p_2 ongelijk zijn. Ze verschillen echter niet zodanig dat dit een groot effect heeft op de ligging van de 2 krommen).

Het verschillende karakter van de neerslag (grootschalig versus convectief) komt vooral tot uiting in het grote verschil tussen de 2 conditionele kansen voor de verschillende stations in de 2 'seizoenen'. Het percentage van de zomergevallen waarin er 's ochtends neerslag valt dat gevolgd wordt door neerslag 's middags, i.e. $P(E_2 \mid E_1)$, is voor Beek veel groter dan de soortgelijke "persistentie" van gistermiddag naar hedenochtend, i.e. $P(E_1 \mid E_2^{-1})$. Ochtendneerslag is hier dus een redelijke indicatie voor ook 's middags neerslag; middagneerslag is dit veel minder voor de neerslag van de volgende morgen. Voor De Bilt is dit veel minder het geval en voor De Kooy geldt zelfs het tegenovergestelde. In het winterhalfjaar vinden we voor alle 3 stations nauwelijks verschillen in "persistenties" tussen

de dagdelen. Het convectieve karakter van neerslag komt "uitvergroot" tot uitdrukking in de conditionele kansen zoals die hier voor opeenvolgende intervallen gedefinieerd zijn.
De covarianties zijn in de 'winter' het grootst.

B. Tenslotte kunnen analoge tabellen gegeven worden voor een opdeling van de dag **in 4 6-uursperioden**. Deze perioden worden aangegeven met indices 1 t/m 4; indices 1 en 2 hebben dus een andere betekenis dan in het bovenbeschreven geval van 2 deelintervallen. De onderstaande tabellen kunnen ook in combinatie met de 2 bovenstaande gebruikt worden (p_1 en p_2 hieronder omvatten een periode gelijk aan die van p_1 boven, etc.). Alleen de frequenties en de conditionele kansen berekend op onze ca 11-jarige dataset zijn gegeven, wederom in procenten.

Zomerhalfjaar

	De Kooy	De Bilt	Beek
p1 (00-06)	18.6	17.3	16.0
p2 (06-12)	17.8	19.6	16.9
p3 (12-18)	16.0	21.3	22.1
p4 (18-24)	17.3	17.7	17.0
p (00-24)	40.1	40.6	39.5
$P(E_1 \mid E_4^{-1})$	51.8	52.8	45.0
$P(E_2 \mid E_1)$	48.9	59.9	57.4
$P(E_3 \mid E_2)$	46.0	55.6	59.9
$P(E_4 \mid E_3)$	50.6	46.5	42.6

Winterhalfjaar

	De Kooy	De Bilt	Beek
p1 (00-06)	24.4	23.3	22.1
p2 (06-12)	22.4	23.0	22.8
p3 (12-18)	23.9	24.5	23.1
p4 (18-24)	24.7	23.8	22.3
p (00-24)	49.8	48.1	46.5
$P(E_1 \mid E_4^{-1})$	57.0	56.3	53.1
$P(E_2 \mid E_1)$	53.9	57.0	59.4
$P(E_3 \mid E_2)$	55.6	58.6	56.3
$P(E_4 \mid E_3)$	58.3	56.1	55.6

Uit de waarden van de 6-uursfrequenties blijkt dat de onder A. gesignaleerde hogere getallen voor de 'middagneerslag' voor de landstations in de 'zomer' grotendeels te wijten zijn aan de 12-18GMT periode. In de 'winter' valt er geen duidelijke dagelijkse gang te bespeuren in de neerslagfrequentie, of het moet de verhoogde kans op nachtelijke (18-06GMT) neerslag zijn in De Kooy. Dit laatste zou mooi overeen komen met het bekende fenomeen dat in het najaar de onweersfrequentie in een smalle strook langs de kust hoger is in

de avond en nacht. Als gevolg van de relatief hoge zeewatertemperaturen ontstaan gemakkelijk buien die boven land snel uitdoven. Zie voor details Können, 1983.

Voor De Bilt is zowel voor de 'zomer' (Z) als de 'winter' (W) het verband tussen de kansen p_1 tot en met p_4 met p weergegeven in de desbetreffende grafiek in paragraaf 4. Hierbij moet wederom bedacht worden dat de curven voor de ondergrens en voor de onafhankelijke situatie bij deze set van ongelijke p_i's een fractie hoger liggen dan in de grafiek is aangegeven.

Er is wederom een grote variatie in de conditionele kansen voor de landstations in de 'zomer', die zich vooral manifesteert als gevolg van de lage waarden voor de kansen op neerslag tussen 18 en 24GMT gegeven het feit dat het regent tussen 12 en 18GMT, i.e. $P(E_4 \mid E_3)$.

7. Slotopmerkingen

In dit pamflet is nader ingegaan op de relatie tussen de neerslagkansen over dagdelen en de neerslagkans over de hele dag. Uiteraard zijn de argumenten en conclusies niet beperkt tot de predictand neerslag, maar gelden voor alle predictands waarvoor kansverwachtingen afgeleid moeten worden. Hetzelfde geldt voor de opdeling van de totale periode in disjuncte deelperioden. Dit hoeven niet noodzakelijkerwijs gelijke perioden te zijn.

Het is tegenwoordig in de weerdienst gebruikelijk om de uitgaande kansverwachtingen te beperken tot de range tussen 10 en 90%. De reden hiervoor is waarschijnlijk dat men wil voorkomen al te "absolute" uitspraken te doen; bij een forecast van 0% kans op neerslag moet men er absoluut zeker van zijn dat het droog blijft. Analoog bij een 100% kansuitspraak. Zo'n zekerheid zal inderdaad niet veel voorkomen maar waarom de grenzen bij 10 en 90 liggen is niet direct duidelijk. De forecaster is wel degelijk in staat om bijv. bij een dag 1 verwachting voor de 24-uurskans op neerslag (PoP24) "skilful" verschillen tussen 5% en 10% kans aan te geven. Met skilful wordt hier bedoeld dat na alle 5% forecasts het gemiddeld minder vaak geregend heeft dan na de 10% forecasts. Een ondergrens van 5% voor de dag 1 neerslagverwachting is dus heel goed te verdedigen. Bij meer zeldzame verschijnselen, zoals bij mist of onweer, is het niet hanteren van een 10% ondergrens natuurlijk evident.

Voor perioden korter dan 24 uur geldt min of meer hetzelfde. Een ondergrens van 10% voor de kans op neerslag in bijv. een 6-uursperiode is klimatologisch gezien een relatief hoge kans (zie ook het getallenvoorbeeld in paragraaf 4). Een 10% kans in een 6-uursperiode en een "analoog" weertype de rest van de dag (bijvoorbeeld een kleine kans op een bui, onafhankelijk van het tijdstip van de dag) resulteert in een 24-uurskans van ca 35% (zie figuur in paragraaf 4). Omgekeerd, als bij zo'n weertype de 24-uurskans op 10% geschat (of berekend) wordt, dan zou de 6-uurskansen in de buurt van de 3% horen te liggen.

Bij het gebruiken van objectieve kansen, bijvoorbeeld verkregen m.b.v. MOS moet men er rekening mee houden dat, tenzij anders ontwikkeld, de kansverwachtingen betrekking hebben op het al of niet vallen van neerslag in het station waarvoor de verwachting geldt. Meer expliciet, eigenlijk alleen maar voor de regenmeter en niet voor een plek een meter verderop. Tenslotte moet, misschien ten overvloede, nog opgemerkt worden dat hoge kansen op neerslag in principe niks zeggen over de hoeveelheid neerslag die verwacht wordt / moet worden; de hele dag motregen kan voor ieder 3-uur tijdvak een grote neerslagkans opleveren, terwijl de totale hoeveelheid op de dag maar zeer weinig is. Hetzelfde geldt andersom: er kan een zeer kleine kans zijn op een zeer zware bui.

Dankbetuiging
Dank is verschuldigd allereerst aan Frank van Lindert die min of meer noodgedwongen de laatste jaren af en toe discussies over kansen aanzwengelt en mij daar deelgenoot van maakt. En ook aan Herman Wessels, Frans Debie, Daan Vogelezang en Hans Hersbach voor de onderhoudende gesprekken over dit onderwerp. Tenslotte aan Rinske Krabbe en Seijo Kruizinga voor het kritisch doorlezen van dit pamflet.

Referenties

Buijs, A., 1997. Statistiek om mee te werken. *Educatieve Partners Nederland*. 416pp.
Krzysztofowicz, R., 1999. Probabilities for a period and its subperiods: theoretical relations for forecasting. *Mon. Wea. Rev.*, **127**, 228-235.
Können, G.P., 1983. Het weer in Nederland. Thieme-Zutphen, 143pp.