K O N I N K L IJ K N E D E R L A N D S
M E T E O R O L O G I S C H I N S T I T U U T

D e B i l t

Verslagen
V - 266

J.P. de Jongh en S. Kruizinga

Enkele onderzoekingen met betrekking tot de maandverwachting

De Bilt, 1975

Publikatienummer: K.N.M.I. V-266 (M.O./S.B.)

U.D.C.: 551.509.331 :
551.509.323

Enkele onderzoekingen met betrekking tot de maandverwachting.

J.P. de Jongh en S. Kruizinga.

Inleiding.

Door de Jongh (zie memorandum MO-A 75-211) en Kruizinga (zie VSB 75/4) is de relatie tussen het maandgemiddelde van de hoogte van resp. het 700mbar en 500mbar vlak en enkele weergrootheden in De Bilt onderzocht. In principe kunnen deze relaties worden gebruikt om op basis van de prognose van het gemiddelde 700mbar vlak van de volgende maand de weergrootheden van de volgende maand te voorspellen. De kwaliteit van de totale voorspelling wordt beïnvloed door de kwaliteit van de prognose van het 700mbar vlak. De kwaliteit van deze prognose is onderzocht door J. van der Tol* (studentenonderzoek). Bij het hier beschreven onderzoek is gebruik gemaakt van de basisgegevens van van der Tol.

Om uit de bovenlucht-gegevens de weergrootheden af te leiden zullen we uitsluitend gebruik maken van de techniek van de lineaire regressie. Hierbij kunnen we nog twee mogelijkheden onderscheiden, namelijk de Perfect Prog (PP) methode en de Model Output Statistics (MOS) methode. Beide methodes komen in dit verslag aan de orde.

Aangezien, zoals in dit verslag zal blijken, de resultaten van de PP-methode en de MOS-methode nogal teleurstellend zijn werden nog enige andere experimenten uitgevoerd, (zie paragraaf 4 en 5).

Notaties.

Omdat de omschrijving van de grootheden waarmee we gaan werken nogal omslachtig is zullen we hier de notaties vastleggen.

T_w : Het waargenomen maandgemiddelde van de temperatuur verminderd met de normaal.

$T_{p,XX}$: De prognose van het maandgemiddelde van de temperatuur verminderd met de normaal, volgens de methode XX.

5 of 7

$h_{w \text{ of } p,n}$: Het waargenomen (of de prognose van het) maandgemiddelde van de hoogte van het 500mbar of 700mbar vlak in roosterpunt n, verminderd met de bijbehorende normaal. Indien de index "w" voorzien is van (-1) dus w(-1) dan heeft dit gegeven betrekking op de voorgaande maand.
Alle hoogtes worden gegeven in decameters.

2. Perfect Prog methode.

Bij de PP-methode volgens de lineaire regressie gaat men als volgt te werk. Men stelt dat de relatie tussen de temperatuur T_w en de hoogten $h^7_{w,n}$ van het 700mbar-vlak in N roosterpunten als volgt kan worden geschreven:

$$T_w = \sum_{n=1}^{N} a_n * h^7_{w,n} + \varepsilon \qquad (2.1)$$

waarin a_n de regressiecoëfficiënten en ε een toevalsvariabele onafhankelijk van alle andere grootheden (voor T_w en $h^7_{w,n}$, zie notaties).
De regressiecoëfficiënten worden zodanig gekozen dat $\bar{\varepsilon}=0$ en σ^2_ε zo klein mogelijk is. De restvariantie σ^2_ε geeft aan hoe goed de regressievergelijking de temperatuur beschrijft.

Bij de PP-methode worden de regressiecoëfficiënten afgeleid uit historisch (geanalyseerd) materiaal. De regressievergelijkingen worden daarna toegepast op de gegevens van de prognose van het 700mbar-vlak.
De standaardfout in de temperatuurprognose is dan opgebouwd uit twee componenten. De eerste component is afkomstig uit de restvariantie; de tweede component is afkomstig uit de fouten in de prognose van het 700mbar-vlak. Onder de aanname dat deze twee componenten onafhankelijk van elkaar zijn zullen we hier een uitdrukking voor de totale fout afleiden.
De voorspelvergelijking is

$$T_{p,PP} = \sum_{n=1}^{N} a_n * h^7_{p,n} \qquad (2.2)$$

De gemiddelde fout in de voorspelde temperatuur is dan

$$\overline{\Delta T} = \overline{(T_{p,PP} - T_w)} \qquad (2.3)$$

$$= \overline{\left(\sum_{n=1}^{N} a_n * h^7_{p,n} - \sum_{n=1}^{N} a_n * h^7_{w,n} - \varepsilon\right)}$$

$$= \sum_{n=1}^{N} a_n * \overline{\left(h^7_{p,n} - h^7_{w,n}\right)} \qquad \text{(want } \bar{\varepsilon} = 0\text{)}$$

De standaarddeviatie van de fout in de temperatuurprognose is:

$$\sigma_{\Delta T} = \sqrt{\overline{\left(T_{p,PP} - T_w - \overline{\Delta T}\right)^2}} \qquad (2.4)$$

De term onder het wortelteken kunnen we ook schrijven als

$$\overline{\left(\sum_{n=1}^{N} a_n * h_{p,n} - \sum_{n=1}^{N} a_n * h_{w,n} - \varepsilon - \overline{\Delta T}\right)^2} \qquad (2.5)$$

Wegens de veronderstelde onafhankelijkheid van regressiefout en prognosefout is dit ook gelijk aan

$$\overline{\left(\sum_{n=1}^{N} a_n * (h_{p,n} - h_{w,n}) - \overline{\Delta T}\right)^2} + \overline{\varepsilon^2} \qquad (2.6)$$

Voeren we nu nog de volgende grootheden in:

$\vec{\vec{C}}_h$: de covariantie matrix van ($\overset{7}{h}_{p,n}$ - $\overset{7}{h}_{w,n}$)

en $\vec{a}$: de N-dimensionale vector $a_1, \cdots a_N$ dan kunnen we (2.6) schrijven als:

$$\vec{a}^T \cdot \vec{\vec{C}}_h \cdot \vec{a} + \sigma_\varepsilon^2 \qquad (2.7)$$

Op grond van het materiaal van v.d. Tol hebben we een schatting gemaakt van deze termen. Ter beschikking stonden 6 jaar (1967-1972) met maandelijks een geanalyseerde kaart en prognosekaart waarvan op 5 roosterpunten de hoogten waren uitgeroosterd. De roosterpunten zijn gekozen op grond van de resultaten van het onderzoek van de Jongh (memo 75-211) en lagen respectievelijk op (20 W, 55 N), (0 W, 70 N), (20 E, 55 N), (0 W, 40 N) en (0 E, 55 N).

De regressieconstanten en covariantie matrix werden berekend op basis van de gegevens over de 5 zomermaanden (mei t/m september) waarbij gewerkt werd met afwijkingen van de normaal om de jaarlijkse gang te elimineren.

De voorspelvergelijking ziet er als volgt uit,

$$T_{p,PP} = -.071 * h^{7}_{p,1} + .009 * h^{7}_{p,2} + .099 * h^{7}_{p,3} + .053 * h^{7}_{p,4} + .060 * h^{7}_{p,5} \qquad (2.8)$$

De restvariantie van de oorspronkelijke regressie is .69 ($^{o}C^{2}$); de totale variantie is 1.06 ($^{o}C^{2}$).

Voor de gemiddelde foutvector in de hoogte van de prognose vonden we:

$$\bar{h} = (-.79, -.38, -1.07, -.13, -1.35)$$

De covariantie matrix (bovendriehoek) van de fout in de hoogte van de prognose is:

28.4	1.1	1.5	2.6	19.0
.05	18.4	2.1	-2.4	4.5
.10	.17	7.9	.7	6.5
.19	-.22	.10	6.5	3.6
.73	.21	.47	.29	24.0

In de onderdriehoek zijn de correlaties tussen de fouten gegeven. Uit deze gegevens volgt dat de gemiddelde fout in de prognose van de temperatuur volgens (2.3) is:

$$\overline{\Delta T} = -.14 \ ^{o}C$$

De bijdrage aan de variantie tengevolge van de fout in de prognose is

$$\vec{a}^{T} * \vec{\vec{C}}_{h} * \vec{a} = .24 \ (^{o}C)^{2}$$

De totale variantie $\sigma^{2}_{\Delta T}$ wordt dus

$$\sigma^{2}_{\Delta T} = .69 + .24 = .93 \ (^{o}C)^{2} \qquad \text{(zie 2.7)}$$

We zien dat de prognose een betrekkelijk geringe bijdrage levert tot de totale fout in de voorspelling. Passen we echter dit prognose systeem toe op de prognoses van het 700mbar-veld van dezelfde maanden waaruit de voorspelvergelijking is afgeleid, dan vinden we een correlatie van -.007 tussen voorspelde en opgetreden temperatuur. Deze lage correlatie geeft aan dat deze methode van voorspellen onbruikbaar is. Bovendien vinden we dan een variantie in de fout van $1.16(^{o}C)^{2}$. Dus een grotere variantie dan de verwachte $.93(^{o}C)^{2}$, dit geeft een indicatie dat er een zekere samenhang is tussen de restfout in de regressie en de fout ten gevolge van de prognose. De aanname omtrent de onafhankelijkheid van deze twee bijdragen aan de fout is dus waarschijnlijk niet terecht. Het verdient aanbeveling dit in de toekomst nader te bestuderen.

3. Model Output Statistics.

Bij deze methode kunnen we ook weer lineaire regressie gebruiken. De regressiecoëfficiënten worden nu echter afgeleid uit historische prognoses met daarbij behorende gemeten temperaturen. Op deze manier is het mogelijk bepaalde (vooral systematische) fouten in de prognose van de 700mbar-hoogten te compenseren. Passen we dit systeem toe op dezelfde maanden als gebruikt in de voorgaande paragraaf, dan vinden we een multiple correlatiecoëfficiënt van .17 tussen de $T_{p,MOS}$ en T_w
De restvariantie (voorspeld minus opgetreden) bedraagt dan $1.05(^{o}C^{2})$ bij een totale variantie van $1.08(^{o}C^{2})$. Dat wil zeggen dat de standaard fout van de prognose van temperatuur t.o.v. de waargenomen temperatuur ongeveer $1^{o}C$ is. Dit is iets beter dan het voorgaande systeem. Bij toepassing van deze MOS techniek blijkt echter dat in dit geval vrijwel altijd de normaal voorspeld wordt waardoor het tamelijk zinloos wordt om dit systeem toe te passen.

4. Verdere experimenten.

Met dezelfde gegevens die voor de hiervoor beschreven experimenten zijn gebruikt is ook nog onderzocht in hoeverre er een verband is tussen $h^{7}_{w(-1),n}$ en T_w.
Ook dit verband is onderzocht in de vorm van een lineaire regressie. De temperatuur-gegevens hadden wederom betrekking op de maanden mei t/m september uit de periode 1967 t/m 1972. We vonden een multiple correlatiecoëfficiënt van .52 en een restvariantie van $.8(^{o}C)^{2}$.

Dit betekent dat de prognose van de temperatuur gebaseerd op deze regressie-vergelijkingen een standaardfout zal hebben van ongeveer 0.9°C. (Dit systeem noemen we de time lag (TL) methode).
Dit is iets beter dan bij de voorgaande procedures. Het aantal gegevens is echter te klein om duidelijke uitspraken te doen over de bruikbaarheid van deze regressievergelijkingen. We hebben dit daarom ook onderzocht voor het 500mbar-vlak waarvoor een groter gegevensbestand aanwezig is.

5. Experimenten met 500mbar-vlak gegevens.

Op het KNMI is een bestand van 500mbar-gegevens aanwezig over het tijdvak 1949 t/m 1970. Deze gegevens zijn gebruikt om een aantal voorspelvergelijkingen van de maandgemiddelde etmaaltemperatuur van de volgende maand af te leiden. De volgende berekeningen zijn uitgevoerd.

5.1 Nagegaan is in hoeverre er een relatie is tussen de maandgemiddelde hoogten van het 500mbar-vlak van de afgelopen maand en de gemiddelde etmaaltemperatuur van deze maand. Hiervan is een vergelijking opgesteld van de volgende gedaante.

$$T_{p,TL} = \sum_{n=1}^{5} a_n * h^{5}_{w(-1),n} \qquad (5.1)$$

In eerste instantie is een voorspelvergelijking voor de maandgemiddelde etmaaltemperatuur van de maanden juni t/m augustus afgeleid. Tabel 5.1 geeft een overzicht van de gevonden relaties tussen T_w en $h^5_{w(-1),n}$.

Tabel 5.1: Resultaten regressie $h^5_{w(-1),n}$ en T_w. Van de regressievergelijkingen zijn temperatuurgegevens van de maanden juni t/m augustus gebruikt en hoogtegegevens van de maanden mei t/m juli.

roosterpunt	reductie variantie(%)	totale reductie variantie(%)
5	16	16
3	1	17
2	1	18
4	0	0
1	0	0

De regressievergelijking ziet er als volgt uit

$$T_{p,TL} = -7.23*10^{-3} * h^{5}_{w(-1),1} + 2.01*10^{-2} * h^{5}_{w(-1),2} + 3.25*10^{-2} * h^{5}_{w(-1),3} - 1.08*10^{-2} * h^{5}_{w(-1),4} + 8.44*10^{-2} * h^{5}_{w(-1),5} \quad (5.1)$$

met h uitgedrukt in decameters en T in graden.

Om een indruk te krijgen wat deze regressievergelijking waard is indien deze als voorspelvergelijking wordt gebruikt, is de PI berekend. Stel het temperatuurinterval wordt in 3 klassen verdeeld, met gelijke frekwenties. Zoals bij de maandverwachting op het KNMI gebruikelijk is, worden 2 klassen steeds ingezet. Dit is op de volgende manier gebeurd: indien de voorspelde grootheid $T_{p,TL}$ groter dan 0 is, dan worden de klassen normaal en boven normaal ingezet; indien de voorspelde grootheid $T_{p,TL}$ kleiner dan 0 is, dan worden de klassen normaal en beneden normaal ingezet. De PI, die met het afhankelijke materiaal op deze manier voor de zomermaanden wordt verkregen is 0.20.

5.2 In het eerste experiment is nagegaan in hoeverre de gemiddelde 500mbar-hoogte van de vorige maand bruikbaar is als predictor van de gemiddelde etmaaltemperatuur van deze maand. Het lijkt ook zinvol om na te gaan in hoeverre de anomalie van de gemiddelde etmaaltemperatuur van de vorige maand ($T_{w(-1)}$) nog een bijdrage tot het voorspellen van de temperatuur van deze maand kan leveren. Naast de hoogten van het 500mbar-vlak werd nu ook deze temperatuur anomalie in de regressie ingevoerd. Tabel 5.II geeft de resultaten; we zien dat $T_{w(-1)}$ geen enkele bijdrage geeft.

Tabel 5.II: Resultaten regressie tussen $h^{6}_{w(-1),n}$, $T_{w(-1)}$ en T_w voor de zomermaanden.

predictor	reductie variantie(%)	totale reductie variantie(%)
roosterpunt 5	16	16
3	1	17
2	1	18
4	0	0
1	0	0
$T_{w(-1)}$	0	0

5.3 Als laatste experiment is bekeken of de voorspelvergelijkingen nog beter worden indien niet de gemiddelde hoogte van de gehele voorafgaande maand als ingangsparameter wordt gebruikt maar de gemiddelde hoogte van de laatste 14 dagen van de vorige maand. Indien we tabel 5.I en 5.III vergelijken, dan blijkt dat dit geen verbetering oplevert. Het resultaat is zelfs iets slechter dan wanneer de 500mbar-gegevens van de gehele vorige maand worden gebruikt maar het verschil is beslist niet significant.

Uit een vergelijking van de resultaten van de 3 experimenten van paragraaf 5 blijkt, dat de eenvoudigste en meest bruikbare voorspelvergelijking in experiment 5.1 wordt verkregen.

Tabel 5.III: Resultaten van de regressie tussen $h^5_{w(-1)}$ en $\overline{T}_w$ waarbij $h^5_{w(-1)}$ betrekking heeft op de laatste 14 dagen van de voorafgaande maand.

roosterpunt	reductie variantie(%)	totale reductie variantie(%)
5	12	12
1	3	15
4	1	16
2	0	0
3	0	0

De PI die met deze voorspelvergelijking op het afhankelijke materiaal wordt verkregen is 0.20. In hoeverre deze PI nog lager wordt op onafhankelijk materiaal is getest op gegevens van de zomermaanden 1971 t/m 1974. Indien dezelfde inzetpolitiek wordt gevolgd als op het afhankelijke materiaal (2 van de 3 klassen worden ingezet), dan blijkt de PI voor de 12 gevallen waaruit het onafhankelijk materiaal bestaat 0.17 te zijn. Dit is in goede overeenstemming met de PI op het afhankelijke materiaal.

6. Conclusies.

Op grond van de experimenten beschreven in paragraaf 2 en 3 kunnen we stellen dat de prognoses van het gemiddelde 700mbar-vlak nog te slecht zijn om hieruit via regressie prognoses te maken voor de gemiddelde temperatuur. Wil dit systeem bruikbaar worden dan zal men eerst de prognose van het 700mbar-vlak drastisch moeten verbeteren. Indien men in de toekomst hieraan wil gaan werken op dit instituut dan wordt door ons aanbevolen om hierbij in plaats van 700mbar-vlak het 500mbar-vlak te gaan voorspellen. Een goede prognose van het 500mbar-vlak zal net zo waardevol zijn als een prognose van het 700mbar-vlak. Het beschikbare gegevensbestand van het 500mbar is echter veel groter.

Uit de experimenten beschreven in paragraaf 4 en 5 volgt dat het waarschijnlijk mogelijk is om een redelijke prognose van de gemiddelde temperatuur te verkrijgen uit het hoogtepatroon van de vorige maand. Dit systeem dient zeker nader onderzocht te worden.
Hierbij dient dan ook onderzocht te worden of andere statistische technieken betere resultaten geven.

7. Slot.

Naar de mening van de schrijvers biedt dit rapport tezamen met de voorgaande rapporten (VSB 75/4 en Memo 75-211) voldoende informatie om een gepland onderzoek t.b.v. de maandverwachting op te zetten.
Er zullen daarom geen verdere inleidende experimenten worden uitgevoerd.

Voor de opzet van een maandverwachtingstechniek kan men twee mogelijkheden onderscheiden:

a. De statistische methode.

Bij deze mogelijkheid werkt men verder in de richting van een time lag methode zoals aangegeven in de paragrafen 4 en 5 van dit rapport. Hierbij dient men te bedenken dat de lineaire regressie niet de enige techniek is die mogelijk bruikbaar is.

b. De fysische methode.

Bij deze methoden komt men tot een prognose op basis van fysisch inzicht en ervaring. De amerikaanse methode is gedeeltelijk hierop gebaseerd. We merken hierbij op dat weliswaar volgens dit rapport de amerikaanse prognose van het 700mbar-vlak vrijwel onbruikbaar is, dit geldt niet ten aanzien van hun temperatuur-prognoses (zie memo MO-A 74-174).

Voor beide methoden geldt dat voor het onderzoek een grote dataset nodig is. Het verdient, zoals reeds eerder is opgemerkt, daarom aanbeveling om uit te gaan van 500mbar-gegevens waarvan reeds een gegevensbestand van 22 jaar aanwezig is. Deze set is mogelijk ook nog te klein; men dient daarom ook na te gaan in hoeverre deze set nog is uit te breiden op basis van grondgegevens.

Het is moeilijk om aan te geven welke methode van verwachtingstechniek bij voorkeur onderzocht dient te worden. Verwacht mag worden dat de statistische methode reeds op korte termijn (enkele jaren) tot bruikbare resultaten kan leiden. Of men hiermee de amerikaanse temperatuur-prognose kan overtreffen is nog niet te voorzien.
De fysische methode kan mogelijk betere resultaten leveren, echter pas op langere termijn.

-o-o-o-o-o-o-o-o-