K O N I N K L I J 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

WETENSCHAPPELIJK RAPPORT

W.R. 75-10

J.P. de Jongh

en

S. Kruizinga

Empirische orthogonale eigenvelden van het 500mbar-vlak.

Een onderzoek naar eigenschappen en bruikbaarheid.

De Bilt, 1975

Publikatienummer: K.N.M.I. W.R. 75-10 (MO)

U.D.C. 551.547.5 :
551.509.314 :
551.509.318 :
551.509.323

Summary

The use of emperial orthogonal functions (EOF) has been recommended by several authors for regression and classification techniques as well as for pattern recognition.
In this report we have studied the properties of the EOF of the 500 mbar level in a region which is useful for the meteorologists in the Netherlands. The applicability of these EOF has been studied by performing two experiments.

The technique of the EOF is known as Principal Component Analysis in statistics. In chapter 2 some standard statistical theory has been given.

The EOF are computed from the daily 500 mbar heights at the grid points indicated in figure 3.1 during the period 1961 till 1971. In chapter 3 the properties of the computed EOF are discussed. Since the mean as well as the variance show a yearly variation the data set was split up in monthly sets from which the EOF were derived. As an example we have given the mean field and the first ten EOF of August in figure 3.2 - 3.12.

Part of the total variance which is explained by the 5, 10 and 20 most important EOF out of 63 is given in table 3.II.

The spatial differences in the contribution to the local variance has been studied in chapter 3.2.
In the figures 3.18 - 3.22 the relative contribution to the local variance of the first five eigenvectors for August has been depicted. From this figures it follows that for August the fifth eigenvector is very important for the Netherlands. In the figures 3.23 - 3.25 the relative local residual variance after applying 5, 10, 20 EOF has been given.

The composition of two actual fields out of the EOF is shown in figures 3.27 - 3.31 and 3.33 - 3.37.
The actual fields belonging to this series are shown in figures 3.26 and 3.32. In chapter 3.3 these composition has been discussed.

The scores on the EOF of two consecutive days are correlated. In the tables 3.IV and 3.V (chapter 3.4) the auto- and crosscorrelations of scores with a time lag of one day are given for August and January. These tables show that especially the autocorrelation of the score of the first EOF (large system) is very high.

The differences between EOF of consecutive months is studied in chapter 3.5. The conclusion is that EOF of successive months are nearly equal. So it is possible to mix the data of several months after proper elimination of the monthly differences in the mean heights.

In chapter 4 two possible applications of the technique of EOF are studied. In the first application the maximum temperature of the next day was predicted as follows. The 500 mbar fields predicted by a numerical model were decomposed in EOF. The maximum temperature belonging to it was computed on the basis of a regression equation. The constants of this equation were derived from decomposed analysed fields and the observed maximum temperatures. It turned out that this procedure performed slightly better than the application of the regression technique on the 500 mbar heights directly.

In the second experiment analogous patterns were selected with the help of EOF as well as with the help of heights directly in several ways. The direct use of the heights corresponded best with the subjective technique used by the meteorologists of our institute.

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

INHOUD

1. Inleiding.

Van de zijde van de operationele weerdienst bestaat grote interesse voor een systeem dat m.b.v. de computer analoge stromingspatronen kan selecteren. Een voorspelmethode waarbij dagelijks analogen worden gezocht wordt momenteel toegepast bij de meerdaagse verwachting (2 en 3 dagen vooruit). De analogen-selectie is echter bijzonder arbeidsintensief. Ook bij het opstellen van de 1-daagse verwachting zal een analogen-selectie een bijdrage kunnen leveren. In de literatuur [1, 2, 3] wordt door een aantal onderzoekers gesuggereerd om bij de analogen-selectie gebruik te maken van empirische orthonormale functies (EOF). Naast het selecteren van analogen bestaat ook veel belangstelling voor het classificeren van prognoses en voor het rechtstreeks afleiden (regressie) van weergrootheden uit prognoses. Deze gegevens kunnen dienen als een eerste indicatie voor de op te stellen verwachting. Ook voor regressie en classificatie wordt het gebruik van EOF door sommige auteurs aanbevolen [4]. Over de practische toepassing is ons geen literatuur bekend zodat de waarde van de EOF niet a priori vast lag. De algemene eigenschappen van de EOF zijn echter zodanig dat het zeker nuttig leek om hun eigenschappen nader te bestuderen en een onderzoek in ge stellen naar de practische toepasbaarheid. In dit rapport is het onderzoek naar deze toepasbaarheid toegespitst op de analogen-selectie en de regressieproblemen.

In hoofdstuk 2 zal iets omtrent de theoretische achtergrond van EOF worden behandeld.

Hoofdstuk 3 geeft een overzicht van een aantal berekende EOF's. Verder zal nog worden ingegaan op de mate waarin de gevonden functies representatief zijn en in hoeverre een bepaalde set van deze functies in staat is het stromingsveld te beschrijven.

In hoofdstuk 4 tenslotte zullen een tweetal voorbeelden van het gebruik van EOF worden gegeven, n.l.

1. De voorspelling van de maximumtemperatuur van de volgende dag m.b.v. de 36 uur-prognose van de BK3.
2. De selectie van analoge stromingspatronen.

De resultaten van verschillende berekeningen zullen worden besproken. Tevens zullen analogen-selectiemethodes worden beschouwd waarbij geen gebruik wordt gemaakt van EOF.

2. Theorie.

De techniek van het ontbinden in empirische orthogonale functies staat in de statistiek bekend onder de naam "principal component analysis". We zullen in dit overzicht de statistische terminologie aanhouden. Het ligt niet in de bedoeling om de volledige theorie te behandelen; hiervoor wordt verwezen naar Dempster [5], Lawley en Maxwell [6] en Kendall en Stuart [7].

In de statistische terminologie wordt het 500 mbar veld van 1 dag aangeduid als een individu waaraan K variabelen (de 500 mbar hoogte op K plaatsen) zijn gemeten. We willen nu trachten de varianties en de kovarianties van de verschillende variabelen zo goed mogelijk samen te vatten in de zogenaamde "principal variables".
Stel we beschikken over een groot aantal N individuen waaraan de waarde van de K variabelen h_k^* is gemeten. Aangezien we slechts geïnteresseerd zijn in de varianties en kovarianties verwijderen we eerst de gemiddelden volgens

$$h_k = (h_k^* - \overline{h_k^*}^{\,N}) \qquad (2.1)$$

(— N = gemiddeld over de individuen)

De verzameling van K hoogte-afwijkingen van het gemiddelde kunnen we beschouwen als een rijvector h die het individu representeert.
De steekproef (ko)variantie matrix S hiervan wordt als volgt gedefinieerd.

$$S = \overline{h^T \circledast h}^{\,N} \qquad (2.2)$$

(⊛ matrix product)
(T = gespiegeld)
(N aantal individuen)

Voor N naar oneindig convergeert deze steekproef kovariantie matrix naar de populatie kovariantie matrix S^*. De eigenwaarden λ^* en eigenvectoren e^* van deze matrix S^* zijn respectievelijk de principal components en de principal variables. De eigenwaarden λ en de eigenvectoren e van de matrix S kunnen worden gezien als schattingen van λ^* en e^*. Het is gewoonte om deze sets λ en e te ordenen naar de grootte

van λ. Dus λ_1, is de grootste eigenwaarde en e_1 de daarbijbehorende eigenvector.

Op grond van de definitie van de eigenvectoren en de eigenschappen van S vormen deze tezamen een orthonormale basis van de K-dimensionale ruimte waarin we de vectoren h kunnen schrijven als lineaire combinatie van de e_k , n.l.

$$h = \sum_{k=1}^{K} \alpha_k \, e_k \tag{2.3}$$

De parameters α_k variëren van individu tot individu en worden de score van het individu op de eigenvector genoemd. Van deze scores kan men de volgende eigenschappen bewijzen:

$$\text{a)} \quad \overline{\alpha_k}^N = 0$$

$$\text{b)} \quad \overline{\alpha_k^2}^N = \lambda_k \tag{2.4}$$

$$\text{c)} \quad \overline{\alpha_k \alpha_j}^N = 0 \qquad k \neq j.$$

Substitueren we (2.3) in (2.2) dan krijgen we

$$S = \sum_{k=1}^{K} \overline{\alpha_k^2}^N \cdot \left(e_k^T \bullet e_k \right) = \sum_{k=1}^{K} \lambda_k \cdot \left(e_k^T \bullet e_k \right) \tag{2.5}$$

Hieruit volgt dat de eerste eigenwaarde en eigenvector het meest bijdragen tot verklaring van de kovariantie matrix. Indien we dus de individuen voortaan zouden karakteriseren door hun score op de eerste eigenvector, hebben we zoveel mogelijk van de variantie en samenhang van het hoogteveld verklaard als op deze manier mogelijk is. Uit het ongecorreleerd zijn van de scores volgt dan bovendien dat de benadering van het hoogteveld door de eerste score maal de eerste eigenvector niet systematisch fout is. Dit zelfde geldt uiteraard voor de benadering

door de twee scores enzovoorts.
De idee die hierachter steekt, is, dat de eerste eigenvelden betrekking zullen hebben op grote meteorologische systemen terwijl de hoger genummerde eigenvelden de zeer locale systemen en de analyse fouten representeren. Door dus de reeks van (2.3) af te kappen na $K' < K$ kunnen we uit ieder veld een nieuw veld reconstrueren dat de globale eigenschappen van het actuele veld weergeeft. Een vaste regel waar we moeten afkappen is niet te geven; deze dient te worden afgeleid uit het materiaal.

Het is belangrijk om op te merken dat de eigenvectoren met de hoogste eigenwaarde weliswaar het belangrijkst zijn voor de beschrijving van de velden doch niet noodzakelijk even belangrijk zijn voor uit het veld afgeleide grootheden. Het volgende voorbeeld illustreert dit verschil. Stel we hebben een kanaal waar zowel aan de ingang als aan de uitgang de waterstand wordt gemeten. Uit een groot aantal metingen berekenen we de volgende kovariantie matrix.

$$S = \begin{bmatrix} 100 & 90 \\ 90 & 100 \end{bmatrix} \quad (m^2) \tag{2.6}$$

dus de hoogtes aan in- en uitgang hebben een standaarddeviatie van 10 meter (variantie = 100 m^2) en een kovariantie van 90 m^2 of een correlatie van $90/\sqrt{100 \times 100} = 0{,}90$.
Voor de eigenwaarden vinden we $\lambda_1 = 190$ en $\lambda_2 = 10$ en voor de eigenvectoren $e_1 = (0.7, 0.7)$ en $e_2 = (0.7, -0.7)$.
Reconstructie volgens (2.5) geeft

$$S = 190 \times \left[\begin{bmatrix} 0.7 \\ 0.7 \end{bmatrix} \circledast (0.7, 0.7) \right] + 10 \left[\begin{bmatrix} 0.7 \\ -0.7 \end{bmatrix} \circledast (0.7, -0.7) \right] =$$

$$= \begin{bmatrix} 95 & 95 \\ 95 & 95 \end{bmatrix} + \begin{bmatrix} 5 & -5 \\ -5 & 5 \end{bmatrix} = \begin{bmatrix} 100 & 90 \\ 90 & 100 \end{bmatrix}. \tag{2.7}$$

We zien dat de matrix S door de eerste eigenvector al heel goed gereconstrueerd wordt. De interpretatie van de eigenvectoren is eenvoudig.

De eerste eigenvector representeert de gemiddelde hoogte en de tweede eigenvector het hoogteverschil. Voor de globale beschrijving van de toestand in het kanaal voldoet de gemiddelde hoogte oftewel de score van de eerste eigenvector zeer goed.
Beschouwen we echter de afgeleide grootheid "de stroomsnelheid in het kanaal" dan zal duidelijk zijn dat voor deze grootheid de score op de eerste eigenvector niet van belang is doch juist de score van de tweede eigenvector.

Zoals hiervoor is gesteld zijn de berekende eigenvectoren en eigenwaarden slechts schattingen van de werkelijke waarden. Belangrijk, voor een toekomstige toepassing op nieuwe individuen, is om te weten hoe goed deze schattingen zijn. Onder de veronderstelling dat h een K-dimensionale normale verdeling heeft, zijn hiervoor formules bekend (zie Lawley en Maxwell [6], pagina 22.) Voor de betrouwbaarheid van de eigenwaarden geldt:

$$\text{var}\,(\lambda_k) = 2\,\lambda^2_k \,/\, N\,. \tag{2.8}$$

Voor de eigenvectoren is geen directe formule gevonden maar wel voor de gewichtsvectoren w_k. Deze zijn gedefinieerd als

$$w_k = \sqrt{\lambda_k}\cdot e_k \tag{2.9}$$

(Zie ook NB1 pag. 8)

Voor de variantie w_{ik} (het i-de element van de k-de vector) geldt:

$$\text{var}\,(w_{ik}) = \left[\frac{w^2_{ik}}{2} + \lambda^2_k \sum_{\substack{e=1\\ e\neq k}}^{K} w^2_{ie}\Big/(\lambda_k-\lambda_e)^2\right]\Big/ N \tag{2.10}$$

en voor de kovariantie tussen w_{ik} en w_{jk} geldt:

$$\operatorname{cov}\left(w_{ik},w_{jk}\right) = \left[\frac{w_{ik}\cdot w_{jk}}{2} + \lambda_k^2 \sum_{\substack{l=1 \\ l\neq k}}^{K} w_{il}\cdot w_{jl} \Big/ \left(\lambda_k - \lambda_l\right)^2\right] \Big/ N \qquad (2.11)$$

We kunnen de toevallige fout in de gewichtsvector opsplitsen in twee delen. Het eerste deel is parallel aan de vector en het tweede deel staat loodrecht op de gewichtsvector. Het eerste deel correspondeert dan met de fout in de eigenwaarde, het tweede deel geeft aan in hoeverre de richting vast ligt. In de formules (2.10) en 2.11) heeft de eerste term van het rechter lid betrekking op het eerste deel, de tweede term van het rechter lid geeft het tweede deel. Nu is het zo dat voor ons gebruik alleen het tweede deel van belang is.
Als de richting van de eigenvector goed vast ligt kunnen we hem ook zinvol op nieuwe hoogtevelden toepassen. Dat dan de variantie van de scores nog tijdsafhankelijk is, is niet zo belangrijk.
De splitsing van de fout in richting en lengte is alleen toegestaan voor zeer grote N. Bij toepassing van de formules bij kleine N moeten de resultaten alleen als indicaties worden gezien en niet als exact.

N.B.1 In de praktijk wordt veel met de gewichtsvectoren gewerkt. Het voordeel is dat het belang van de vector meteen in zijn lengte tot uitdrukking komt. Uiteraard blijft (2.3) geldig alleen de eigenschappen van de α's worden:

$$\overline{\alpha_k}^{N} = 0$$

$$\overline{\alpha_k^2}^{N} = 1 \qquad (2.4)'$$

$$\overline{\alpha_k \alpha_j}^{N} = 0 \quad k \neq j$$

en (2.5) wordt:

$$S = \sum_{k=1}^{K} \left(w_k^T \circledast w_k \right) . \qquad (2.5)'$$

N.B.2 Het is van belang om te memoreren dat de eigenvectoren slechts de variantie en de samenhang tussen de roosterpunten beschrijven en in feite niet de ruimtelijke samenhang van de meteorologische velden. Dit uit zich vooral in het volgende: indien we bij een andere studie een andere roosterpuntenconfiguratie kiezen zullen we zowel andere velden als andere volgorde van belangrijkheid kunnen vinden. Dit effect zal zich waarschijnlijk vooral uiten in de grootte van de eigenwaarden en niet zozeer in de verandering van de bij de eigenvectoren horende velden. De in dit rapport gepubliceerde velden zijn daarom in principe gebonden aan het gebruikte rooster. We zullen dit later nog experimenteel onderzoeken.

3. Overzicht en eigenschappen van de eigenvelden.

3.1 Beschrijving eigenvelden.

Op een bestand van 500 mbar gegevens zijn empirische orthogonale eigenfuncties berekend. Dit bestand bevat de hoogtes van het 500 mbar-vlak om 12.00 z in een aantal roosterpunten (zie fig. 3.1) gedurende het tijdvak 1961 t/m 1970.

Daar zowel het gemiddelde, de variantie en ruimtelijke correlatie van de 500 mbar hoogtes een jaarlijkse gang vertonen werd het materiaal naar maand opgesplitst. De geinterpoleerde velden behorend bij de eigenvectoren zullen we eigenvelden noemen.

De eigenvectoren zijn in eerste instantie berekend op een veld van 63 roosterpunten zoals dat in fig. 3.1 is aangegeven. Het kaartgebied met de 63 roosterpunten zal in het vervolg het "grote veld" worden genoemd. Een gedeelte van de berekeningen is herhaald op een deelgebied van het grote veld, zoals in fig. 3.1 is aangegeven ("klein veld").

Voor iedere maand zijn de eigenvelden berekend; in dit rapport zullen slechts een aantal voorbeelden worden behandeld.

In overeenstemming met de theorie zijn eerst de gemiddelde hoogtevelden opgemaakt. Als voorbeeld wordt in fig. 3.2 het gemiddelde hoogteveld van augustus gegeven. In de figuren 3.3 t/m 3.12 zijn de eerste tien gewichtsvelden van augustus afgebeeld. De velden zijn dus zodanig genormeerd dat $\sum_{i=1}^{63} w_{ik}^2 = \lambda_k$; dit komt dus overeen met de gewichtsvectoren. Voor de figuren zijn de gewichtsvelden gebruikt ten gerieve van de aanschouwelijkheid, normaal werken we met de eigenvelden.

Aan de eerste paar eigenvelden is nog een zekere meteorologische interpretatie te geven, maar de eigenvelden met hoger rangnummer worden al spoedig te gecompliceerd.

De eerste eigenveld van augustus (fig. 3.3) geeft op de Atlantische Oceaan een O-W stroming indien de score positief is; indien de score negatief is zal er op de gemiddelde kaart een W-O stroming worden gesuperponeerd. Indien de score van het tweede eigenveld

positief is, dan zal een N-Z stroming boven de Atlantische Oceaan bij de gemiddelde kaart worden opgeteld.
Bij verdere pogingen tot interpretatie van de gegeven kaarten dient bedacht te worden dat de kaarten een samenvatting geven van de statistische eigenschappen van het hoogteveld welke niet noodzakelijk samenvallen met meteorologisch karakteristieke patronen.

Tabel 3.1 geeft een overzicht van de eerste 20 eigenwaarden die op het materiaal van de augustus-maanden zijn gevonden. Tevens is de bijdrage aan de variantie opgegeven die de verschillende eigenvectoren leveren. De bijdrage van de i^e eigenvector aan de varianties is gedefinieerd als:

$$\frac{\lambda_i}{\sum_{i=i}^{63} \lambda_i} .$$

Tevens wordt in deze tabel de totale bijdrage aan de variantie van deze eigenvector plus alle voorgaande eigenvectoren gegeven.

Tabel 3.1: Eigenwaarden van de eigenvector 1 t/m 20 en de bijdrage tot de reductie van de variantie van de bijbehorende eigenvectoren.

n	λ_n	bijdrage var (%)	totale bijdrage var (%)	n	λ_n	bijdrage var (%)	totale bijdrage (var (%)
1	1432	27.4	27.4	11	80.0	1.5	90.7
2	790	15.1	42.5	12	66.8	1.3	92.0
3	643	12.3	54.8	13	63.3	1.2	93.2
4	505	9.7	64.4	14	47.2	0.9	94.1
5	376	7.3	71.6	15	42.0	0.8	94.9
6	248	4.7	76.3	16	34.2	0.7	95.6
7	223	4.3	80.6	17	28.2	0.5	96.1
8	164	3.1	83.7	18	22.3	0.4	96.6
9	154	2.9	86.7	19	19.2	0.4	96.9
10	132	2.5	89.2	20	17.1	0.3	97.3

3.2 Ruimtelijke verschillen in de bijdrage aan de variantie.

Van de eerste eigenvelden van de maand augustus is nagegaan in hoeverre ze op de verschillende roosterpunten bijdragen tot de variantie. Het is duidelijk dat de bijdrage tot de variantie 0 is op plaatsen waar de eigenvelden 0 zijn, de bijdrage tot de variantie zal maximaal zijn in de omgeving van de gebieden waar de eigenvelden een extreme waarde hebben.

Het eerste eigenveld van de maand augustus zal meer dan 80% van de variantie verklaren in een gebied ten westen van Noorwegen (fig. 3.18). In de omgeving van Nederland is de bijdrage tot de variantie van het veld door het eerste eigenveld nog slechts 20%, terwijl in het zuidelijk gedeelte van het kaartgebied het eerste eigenveld nauwelijks enige bijdrage levert tot de verklaring van het veld.

Het derde en vierde eigenveld (fig. 3.20 en 3.21) leveren bijna geen bijdrage tot de variantie in de omgeving van Nederland in tegenstelling tot het vijfde eigenveld dat bijzonder essentieel is voor de beschrijving van het veld in onze omgeving; bijdrage tot de variantie ongeveer 35% (zie fig. 3.22).

In fig. 3.23 is de relatieve rest-variantie aangegeven nadat het veld is samengesteld uit de eerste vijf eigenvelden. Er zijn gebieden op de kaart die bijna volledig kunnen worden beschreven met de eerste vijf eigenvelden. In de omgeving van 65° N, 10° O is de restvariantie slechts 7% (16 dam^2 oftewel een standaarddeviatie van 4 dam). Het toevoegen van de eigenvelden 6 t/m 63 zal dus nauwelijks verbetering geven van de beschrijving van het veld daar ter plaatse. Aan de zuidelijke rand van het kaartgebied is de restvariantie nog zeer groot. In de omgeving van Spanje wordt door de eerste vijf eigenvelden slechts 20% van de variantie in het veld verklaard. Uit fig. 3.24 blijkt dat na het toevoegen van de eerste tien eigenvelden de restvariantie aan de zuidrand van het kaartgebied al is teruggebracht tot ongeveer 40%.

Na het toevoegen van 20 eigenvelden (fig. 3.25) is de restvariantie nagenoeg overal kleiner dan 10%; op de meeste plaatsen zelfs minder dan 5%.

In tabel 3.II is de bijdrage van de variantie gegeven na het toevoegen van 5, 10 en 20 eigenvectoren voor de 12 maanden. Er is duidelijk een jaarlijkse gang in de bijdrage aan de variantie. Voor het beschrijven van de stroming in de zomer zijn meer eigenvectoren nodig dan in de winter om dezelfde nauwkeurigheid te behalen.

Tabel 3.II: Bijdrage aan de variantie na het toevoegen van 5, 10 en 20 eigenvectoren per maand.

aantal eigen vect.	j	f	m	a	m	j	j	a	s	o	n	d
5	77	82	79	75	75	71	72	72	75	77	78	79
10	92	94	92	91	91	89	89	89	91	92	92	93
15	98	99	98	98	98	97	97	97	98	98	98	98

Zoals aan het einde van hoofdstuk 2 al staat vermeld, zal de onderlinge positie van de roosterpunten van invloed zijn op de berekende eigenvelden. Om na te gaan hoe groot deze invloed is, zijn er ook eigenvelden berekend op het grote veld waarbij de roosterpunten die in fig. 3.I zijn aangegeven met de cijfers 1 t/m 8, zijn weggelaten. Op deze wijze blijft er een veld van 55 roosterpunten over. In de figuren 3.13 t/m 3.17 zijn de eerste vijf eigenvelden op het rooster van 55 punten getekend. De velden zijn weer zodanig genormeerd dat $\sum_{i=1}^{55} w_{ij}^2 = \lambda_j$.

Een vergelijking tussen de figuren 3.3 t/m 3.7 en 3.13 t/m 3.17 laat zien, dat voor de eigenvelden 1, 2, 3 en 5 de overeenkomst bijzonder groot is; voor het vierde eigenveld blijken er wel enige verschillen te zijn hoewel ook daar de grote lijnen in de patronen wel dezelfde zijn (tekenverschil van corresponderende velden is niet essentieel).
De conclusie lijkt dan ook wel gewettigd, dat het weglaten van een aantal roosterpunten weinig invloed heeft op de vorm van de eerste eigenvelden.

Uit het voorafgaande is wel duidelijk geworden, dat het aantal eigenvelden, dat nodig is om het veld met een bepaalde nauwkeurigheid te beschrijven sterk afhankelijk is van de plaats.

3.3 Samenstellen van het 500 mbar patroon uit de componenten.

Om een indruk te krijgen hoe een 500 mbar patroon wordt opgebouwd door de verschillende componenten, zijn voor twee gevallen de stromingspatronen getekend na het achtereenvolgens toevoegen van de verschillende eigenvelden. Het eerste geval betreft het 500 mbar patroon op 24-8-62, 1200z. Het patroon wordt gekenmerkt door de straalstroom in de buurt van de 50e breedtegraad. Tabel 3.III geeft een overzicht van de scores van de eerste tien eigenvelden.

Tabel 3.III: Scores van de 500 mbar patronen 12.00z van 24-8-62 en 12-8-64.

eigen vector	1	2	3	4	5	6	7	8	9	10
scores 24-8-62	-69	-36	+11	-10	-16	+11	-2	+10	-11	-1
scores 12-8-64	+96	+38	-32	+20	-5	+14	-7	-3	-3	-5

De score van het eerste eigenveld is negatief. Uit fig. 3.3 blijkt dat een negatieve score van het eerste eigenveld de zonale stroming die in het gemiddelde veld al aanwezig is (fig. 3.2), zal versterken. Het toevoegen van de tweede component zal door zijn negatieve score vooral op het westelijk gedeelte van het kaartgebied een versterking van de straalstroom geven. Na het toevoegen van 10 componenten lijkt de dan ontstane kaart (fig. 3.31) heel redelijk op de analyse (fig. 3.26), hoewel de samengestelde kaart een iets te "glad" verloop heeft. De figuren 3.27 t/m 3.30 geven een aantal tussentoestanden.

Als tweede voorbeeld is het stromingspatroon op 12-8-64, 12.00z gekozen. Dit hoogteveld wordt gekarakteriseerd door de geringe hoogteverschillen (fig. 3.32). Uit tabel 3.III blijkt, dat de score van het eerste eigenveld sterk positief is.

Wordt de eerste component bij het gemiddelde veld opgeteld, dan zal de zonale stroming die op het gemiddelde veld aanwezig is, sterk worden verminderd (fig. 3.33). Na het toevoegen van 10 componenten (fig. 3.37) blijkt dat het dan ontstane patroon goed gelijkt op de analyse van 12-8-64. Wederom geven de kaarten 3.34 t/m 3.36 een aantal tussentoestanden.

3.4 Persistentie.

Het is een welbekend feit dat het 500 mbar hoogteveld nogal persistent is van dag op dag. Op grond hiervan kunnen we verwachten dat de scores op de eigenvelden van dag op dag ook persistent zijn. We hebben dit onderzocht door de autocorrelatie-coëfficiënt van de tijdreeksen van scores te berekenen. Tevens zou men zich kunnen voorstellen dat een hoge score op het eerste eigenveld op de eerste dag aanleiding geeft tot een hoge score op een ander eigenveld op de tweede dag. Daarom zijn ook de kruiscorrelaties tussen de verschillende scores met één dag verschuiving berekend.

In tabel 3.IV zijn de resultaten van deze berekening voor de maand augustus vastgelegd en in tabel 3.V de resultaten voor januari. Uit deze tabellen blijkt dat vooral de scores van de eerste eigenvelden erg persistent zijn. Meteorologisch gezien wil dit zeggen dat vooral, zoals was te verwachten, de grootschalige systemen erg persistent zijn. De eigenvelden met een hoger rangnummer, die samenhangen met kleinschaliger systemen van het drukpatroon, zijn duidelijk minder persistent.

Bij het beschouwen van de kruiscorrelaties dient men te bedenken dat de verschillende scores gemeten op één en dezelfde dag ongecorreleerd zijn. Dus de kruiscorrelaties kunnen niet het gevolg van persistentie zijn. Een duidelijke uitspraak of de gegeven kruiscorrelaties significant zijn kan op dit moment niet gegeven worden. Wel valt op dat zowel in augustus als in januari de hoge kruiscorrelaties duidelijk te vinden zijn bij de kleinere systemen. Dit zou erop wijzen dat kleinere systemen wel in elkaar overgaan op de termijn van een dag; de grote systemen echter niet.

Het is echter ook mogelijk dat bij de grote systemen niet lineaire processen een rol spelen bij de overgangen; ook in dat geval is correlatie nul mogelijk.

Tabel 3.IV: Auto- en kruiscorrelaties tussen de scores van augustus, verschuiving één dag.

Score dag 1 \ Score dag 0	1	2	3	4	5	6	7	8	9	10
1	+.93	-.09	+.01	-.01	-.01	-.01	-.04	-.01	-.02	+.09
2	+.10	+.87	+.01	-.16	+.06	+.05	-.10	-.01	+.02	-.05
3	-.02	-.04	+.90	-.08	+.03	-.02	+.01	+.03	+.07	+.01
4	+.02	+.13	+.06	+.83	-.09	-.12	+.01	+.09	-.15	-.08
5	+.00	-.01	-.02	+.09	+.82	-.05	-.08	+.06	+.03	+.16
6	-.00	+.02	+.11	+.15	-.10	+.74	-.07	+.12	+.13	+.13
7	+.01	+.10	+.02	-.05	+.10	+.06	+.78	-.03	-.06	+.14
8	-.02	-.09	-.09	-.20	+.10	+.09	+.03	+.71	-.13	-.09
9	+.03	+.14	+.12	+.06	-.11	-.29	+.07	+.31	+.61	-.09
10	-.11	+.03	-.04	-.14	-.26	-.20	-.14	+.18	-.12	+.53

Tabel 3.V: Auto- en kruiscorrelaties tussen de scores van januari, verschuiving één dag.

Score dag 1 \ Score dag 0	1	2	3	4	5	6	7	8	9	10
1	+.93	-.07	+.00	-.05	-.03	+.07	-.00	-.05	-.11	-.01
2	+.03	+.90	-.01	-.14	-.01	-.05	+.02	-.10	+.05	+.09
3	+.02	-.02	+.92	-.04	+.05	+.10	-.02	-.04	+.07	+.04
4	+.01	+.19	+.02	+.75	+.06	+.11	+.12	+.03	+.07	-.21
5	-.00	+.04	-.01	-.00	+.82	-.16	-.02	+.01	-.03	+.03
6	-.04	+.03	-.02	-.17	+.18	+.67	-.12	+.07	+.01	-.03
7	-.06	-.09	-.01	-.26	+.06	-.07	+.71	-.12	+.05	-.08
8	-.00	+.04	+.06	-.12	-.04	+.02	-.00	+.73	-.04	-.07
9	+.12	-.00	-.08	+.03	+.06	-.21	-.12	+.09	+.58	+.15
10	+.02	-.01	-.00	+.07	+.08	+.24	+.30	+.03	+.20	+.37

Uit de zeer hoge autocorrelatie van de score van eigenveld 1 volgt tevens dat we in feite slechts over weinig onafhankelijke waarnemingen beschikken. Dit is vooral belangrijk bij de vraag hoe nauwkeurig de eigenvectoren zijn te berekenen. De formules die hiervoor worden gegeven berusten op de veronderstelling dat de waarnemingen onafhankelijk zijn in de tijd. Indien we aannemen dat de reeks een Markov -1 reeks vormt dan kunnen we het aantal effectieve waarnemingen per maand afschatten met de formule van Bartels [8].

$$N_{eff} = \left[\frac{(1 + p1)}{(1 - p1)} - 2. \frac{p1}{n} \cdot \frac{(1-p1)^n}{(1-p1)^2} \right]^{-1} . n \qquad (3.2)$$

waarin p1 de autocorrelatie van dag op dag en n het aantal dagen in de maand. We vinden in ons geval Neff $\approx$ n/14 $\approx$ 2,5 waarnemingen per maand.

In totaal hebben we dan tien maanden dus ongeveer 25 onafhankelijke waarnemingen. In dit effectieve aantal is alleen de autocorrelatie van het eerste eigenveld verwerkt. Aangezien de autocorrelaties van de overige scores lager zijn zal het effectieve aantal dagen vermoedelijk wel hoger zijn dan de genoemde 2,5 per maand. In dit rapport zullen we uitgaan van ongeveer vijf onafhankelijke dagen per maand. Dus ongeveer vijftig onafhankelijke dagen totaal.

3.5 Verschillen tussen de eigenvectoren van de opeenvolgende maanden.

Het is van belang om na te gaan of de eigenvectoren van twee opeenvolgende maanden voldoende verschil vertonen zodat een scheiding naar maanden zinvol is. Mocht dit namelijk niet zo zijn dan zou men twee opeenvolgende maanden bij elkaar kunnen voegen en zo een geringer aantal basisvelden overhouden welke bovendien uit meer materiaal zijn afgeleid. Het probleem hierbij is welke eigenvectoren moeten we onderling vergelijken. Indien we zonder meer steeds de eerste eigenvector met de andere eerste eigenvector

vergelijken zullen we zeker grote verschillen vinden daar het belang van de "eerste" eigenvector wel eens zodanig kan dalen dat het bij een andere maand een "tweede" eigenvector wordt. We kunnen daarom niet zonder meer de eerste eigenvectoren vergelijken. We hebben daarom ook bij iedere eigenvector de best passende eigenvector van de daaraan voorafgaande maand gezocht. Als het criterium voor het gelijk zijn van twee eigenvectoren is het inproduct van de twee eigenvectoren gebruikt. Hierbij dient opgemerkt te worden dat teken omkeer bij eigenvectoren niet ter zake doet zodat ook negatieve inproducten van belang zijn. Verder zijn de eigenvectoren genormeerd op lengte 1 zodat +1 of -1 het maximaal te verwachten inproduct is. In tabel 3.VI zijn de resultaten van de vergelijking voor respectievelijk eigenvectoren 1 en 2 samengevat.

Tabel 3.VI: Vergelijking van de eerste en tweede vector met eigenvectoren van de daaraan voorafgaande maand.

In kolom 1 en 4, inproduct met vector met hetzelfde rangnummer;
In kolom 2 en 5, rangnummer van de best passende eigenvector;
In kolom 3 en 6, inproduct met de best passende eigenvector.

	Eivec. 1			Eivec. 2		
	1	2	3	4	5	6
Januari	.91	1	.91	.94	2	.94
Februari	.93	1	.93	.91	2	.91
Maart	.82	1	.82	-.83	2	-.83
April	.00	2	-.96	-.11	1	-.72
Mei	.82	1	.82	.77	2	.77
Juni	-.75	1	-.75	-.32	3	-.84
Juli	-.13	3	-.66	-.09	1	+.91
Augustus	-.22	2	.94	.15	1	.87
September	-.68	1	-.68	.44	4	.55
Oktober	-.38	2	-.64	.41	1	-.86
November	-.89	1	-.89	-.69	2	-.69
December	-.85	1	-.85	-.80	2	-.80

Bij de beoordeling van deze inproducten dienen we rekening te houden met de statistische schattingsfout.

Volgens formule (2.10) is de variantie van de verschilvector loodrecht op de gewichtsvector gelijk aan:

$$\text{var } (\Delta w) = \sum_{i=1}^{K} \left[\lambda_k^2 \sum_{\substack{l=1 \\ l \neq k}}^{K} w_{il}^2 \Big/ (\lambda_k - \lambda_l)^2 \right] \Big/ N \qquad (3.3)$$

waarin N het aantal onafhankelijke waarnemingen. Verder geldt voor de fout in de eigenvector:

$$\text{var } (\Delta e_k) = \text{var } (\Delta w_k) / \lambda_k \qquad (3.4)$$

waarbij we de fout in λ_k verwaarlozen. Passen we (3.3) en (3.4) toe op de eigenvectoren van augustus dan vinden we dat
var (Δe_1) = .137 waarbij we verondersteld hebben N = 50. Hieruit volgt voor de verwachtingswaarde van het inproduct van het verschil van twee onafhankelijk bepaalde eigenvectoren

$$E\,[(e_1 - e_1^*) \circledast (e_1 - e_1^*)] = \text{var } (\Delta e_1) = .137$$

of

$$E\,[(e_1 \circledast e_1) + (e_1^* \circledast e_1^*) - 2\,(e_1 \circledast e_1^*)] = .137 \qquad (\circledast = \text{inproduct})$$

$$\quad \| \qquad \| \qquad \|$$
$$\quad 1 \qquad 1$$

dus

$$E\,(e_1 \circledast e_1^*) = .93.$$

Op grond hiervan zouden we verwachten dat het inproduct van de twee eigenvectoren van opeenvolgende maanden zo rond de 0.93 zou liggen indien deze eigenvectoren in werkelijkheid hetzelfde waren.

Passen we dezelfde berekeningen toe op augustus de tweede eigenvector en op januari de eerste en tweede eigenvector dan vinden we resp. 0.64, 0.81 en 0.73. De algemene indruk is dat de eigenvectoren van opeenvolgende maanden als niet verschillend beschouwd moeten worden. Het heeft weinig zin om uitvoeriger hieraan te rekenen daar toch geen exacte test mogelijk is. Ten eerste omdat het aantal onafhankelijke waarnemingen onbekend is en ten tweede wegens het benaderende karakter der formules (3.3) en (3.4). Het verdient aanbeveling om bij volgende onderzoekingen de mogelijkheid van een splitsing naar seizoenen na te gaan.

4. Toepassingen.

4.1 Inleiding.

De techniek van het ontbinden van het 500 mbar-vlak in empirische orthogonale functies (eigenvelden) is op een aantal manieren toepasbaar bij het opstellen van de weersverwachting. Twee mogelijke toepassingen zijn nader onderzocht, n.l.:

1. Schatting van de maximumtemperatuur (T_x) van de volgende dag m.b.v. berekende scores en BK3-prognose.
2. Selecteren van analoge stromingspatronen.

4.2 Schatting van T_x m.b.v. scores.

1. Bij de bepaling van regressievergelijkingen maken we gebruik van de zogenaamde "Perfect Prog" methode [9]. Bij deze methode worden regressievergelijkingen afgeleid uit waargenomen materiaal welke daarna worden toegepast op prognoses; in dit geval is dat de 36-uurs BK3-prognose van het 500 mbar hoogteveld

Het is te verwachten dat er een verband bestaat tussen de scores en meteorologische parameters zoals b.v. de maximumtemperatuur, de minimumtemperatuur, etc. Indien b.v. in augustus de score van het eerste eigenveld sterk positief is, dan zal een O-wind relatief warme continentale lucht aanvoeren (zie fig. 3.3); is de score van het eerste eigenveld sterk negatief dan zal met een W-wind in het algemeen koudere oceaanlucht worden aangevoerd. Het lijkt daarom ook zinvol om m.b.v. multiple regressie het verband tussen de scores en T_x nader te onderzoeken. Er is een relatie gezocht in de vorm van

$$T_x = \overline{T}_x + \sum_{i=1}^{P} c_i \cdot \alpha_i$$

waarin: $\overline{T}_x$: gemiddelde maximumtemperatuur in een bepaalde maand.
α_i : score van het i^e eigenveld.
c_i : regressie-coëfficiënten.

Met behulp van waarnemingsmateriaal uit de periode 1961 t/m 1970 is een regressievergelijking voor de maximumtemperatuur in januari en augustus te De Bilt bepaald. De methode waarop de coëfficiënten c_i worden berekend is beschreven door Hamaker [10].

Omdat het materiaal naar maanden is opgesplitst, zijn de regressievergelijkingen op 310 gevallen afgeleid. Gezien de hoge graad van persistentie van twee opeenvolgende dagen zowel wat betreft de maximumtemperatuur als het 500 mbar stromingspatroon, zijn deze 310 gevallen beslist niet als onafhankelijk te beschouwen; de 310 gevallen zullen misschien equivalent zijn met 50 onafhankelijke gevallen.

De regressievergelijking met de eerste vijf scores van T_x in de maand augustus ziet er als volgt uit:

$$T_x = 20.8 + 0.040\ \alpha_1 - 0.045\ \alpha_2 - 0.017\ \alpha_3 - 0.035\ \alpha_4 + \\ + 0.070\ \alpha_5.$$

Zoals al viel te verwachten, is de coëfficiënt van de eerste score positief. Verder is het opmerkelijk dat het vijfde eigenveld belangrijk is zoals uit de grootte van de coëfficiënt blijkt.

Tabel 4.I: Resultaten van regressie van T_x op scores.

januari

variabele	reductie variantie in %	totale reductie variantie in %
α 1	26.0	26.0
α 5	3.6	29.6
α 4	2.3	31.9
α 2	0.2	32.1
α 3	0.1	32.2

augustus

variabele	reductie variantie in %	totale reductie variantie in %
α 1	20.2	20.2
α 5	16.1	36.3
α 2	13.8	50.1
α 4	5.4	55.5
α 3	1.6	57.1

In tabel 4.I wordt een overzicht gegeven van de bijdrage van de eerste vijf componenten tot het berekenen van T_x. Het blijkt, dat het berekenen van T_x uit de scores in januari niet zinvol is; slechts 32% van de variantie in T_x kan met de eerste vijf scores worden verklaard. Aangezien de standaarddeviatie van T_x in januari 4.3°C is, zal de standaardfout in de berekende T_x dan 3.5°C zijn. Indien niet vijf maar twintig scores in de regressievergelijking worden meegenomen, dan blijkt dat de voorspelling van T_x slechts een zeer geringe fractie nauwkeuriger wordt.

De berekening van T_x in de maand augustus op grond van de scores gaat aanmerkelijk beter dan in januari zoals uit tabel 4.I blijkt. M.b.v. de scores van de eerste vijf componenten is het mogelijk om 57% van de variantie van T_x te verklaren. De standaarddeviatie van T_x in de maand augustus is 3.4°C; de standaardfout in de berekende T_x is derhalve 2.2°C.

Er zijn ook regressievergelijkingen afgeleid waarbij de maximumtemperatuur van de "vorige dag" $T_{x,-1}$ (de dag voorafgaande aan de dag waarvoor T_x berekend wordt) wordt gebruikt.
Tabel 4.II geeft een overzicht van de resultaten van de opgestelde regressievergelijkingen.

Tabel 4.II: Resultaten van regressie van T_x op scores en $T_{x,-1}$.

januari

variabele	reductie variantie in %	totale reductie variantie in %
$T_{x,-1}$	63.9	63.9
α 1	1.9	65.8
α 5	1.5	66.3
α 4	0.4	66.7
α 3	0.0	66.7
α 2	0.0	66.7

augustus

variabele	reductie variantie in %	totale reductie variantie in %
$T_{x,-1}$	44.7	44.7
α 5	3.9	48.6
α 2	4.3	52.9
α 1	5.7	58.6
α 4	1.3	59.9
α 3	1.1	61.0

Het blijkt dat in januari $T_{x,-1}$ een bijzonder goede ingangsparameter is (de persistentie is in deze maand kennelijk bijzonder groot) terwijl de bijdrage van de scores uiterst gering is. Bij de voorspelling van T_x in de maand augustus blijken de scores wel degelijk een bijdrage te leveren tot de voorspelling van T_x. In de praktijk lijkt een regressievergelijking waarbij $T_{x,-1}$ ingangsparameter is, wel zinvol te zijn. Op het moment dat de verwachting voor de maximumtemperatuur voor de volgende dag wordt opgesteld (11.00 GMT) is al een goede schatting te maken van $T_{x,-1}$.

2. Toepassing regressievergelijkingen op BK3-prognoses.

Met behulp van BK3-prognoses en analyses uit de maanden augustus en september 1974 is nagegaan in hoeverre de opgestelde regressievergelijkingen bruikbaar zijn. Om na te gaan in hoeverre fouten in de BK3-prognoses invloed hebben op de voorspelde waarde van T_x is de regressievergelijking zowel toegepast op de 36-uur prognose van de 0-uur serie als op de 12-uur-analyse van de dag volgend op de uitgangsdag (prognose en analyse hebben dus betrekking op hetzelfde tijdstip).

In tabel 4.III worden resultaten van de temperatuurvoorspelling gegeven, opgesplitst naar de maanden augustus en september.

Tabel 4.III: Resultaten van schatting van T_x uit BK3-prognoses en analyses (scores) en van weerkamerschattingen. Som van de absolute fouten in ^oC.

			Regressievergelijkingen met $T_{x,-1}$		Regressievergelijkingen zonder $T_{x,-1}$	
	aantal gevallen	$\Sigma\|T_o - T_c\|$ weerkamer	$\Sigma\|T_o - T_c\|$ analyse	$\Sigma\|T_o - T_c\|$ 36h prog	$\Sigma\|T_o - T_c\|$ analyse	$\Sigma\|T_o - T_c\|$ 36h prog
aug.	23	30	37	36	59	58
sept.	20	23	21	30	24	34
totaal	43	53	58	66	83	92

Als maat voor de nauwkeurigheid van de temperatuurverwachting is de som van de absolute waarde van het verschil tussen de opgetreden maximumtemperatuur T_o en de berekende maximumtemperatuur T_c genomen.

De verwachtingen die zijn opgesteld door de weerkamermeteoroloog zijn het nauwkeurigst; totaal van de afwijkingen is 53°C. Verder blijken de temperatuurverwachtingen waarbij $T_{x,-1}$ is gebruikt aanmerkelijk nauwkeuriger te zijn dan de verwachtingen zonder gebruik van $T_{x,-1}$.

Indien $T_{x,-1}$ wordt gebruikt en er geen fouten in de 36h BK3-prognose zouden zijn, dan is de temperatuurverwachting m.b.v. de regressievergelijking slechts een fractie onnauwkeuriger dan de verwachting van de weerkamermeteorologen ($\Sigma|T_o - T_c| = 58^o$).

Door fouten in de BK3-prognoses worden de temperatuurverwachtingen iets onnauwkeuriger zoals uit tabel 4.III blijkt. Indien geen gebruik wordt gemaakt van $T_{x,-1}$ dan blijken de verwachtingen zowel op de analyse als op de prognose veel onnauwkeuriger dan de weerkamerverwachting uit te vallen.

In tabel 4.IV is aangegeven wat de nauwkeurigheid van de verwachtingen zou zijn in het geval van een persistentie-verwachting.

Tabel 4.IV: Resultaten van persistentie en regressie-schattingen uit $T_{x,-1}$. (T_o: gemeten temperatuur, T_c: berekende temperatuur.)

	aantal gevallen	$\Sigma\|T_o - T_c\|$ persistentie	$\Sigma\|T_o - T_c\|$ regressie
aug.	23	51	46
sept.	20	38	31
totaal	43	89	77

Met een totale afwijking van 89°C blijkt dit geen zinvolle verwachtingsmethode te zijn. Tevens is in de tabel aangegeven wat de totale afwijking in de verwachtingen is indien geen scores worden gebruikt, d.w.z. een regressievergelijking in de vorm van

$$T_x = \overline{T}_x + C.(T_{x,-1} - \overline{T}_x)$$

De temperatuurverwachting die op deze manier worden verkregen zijn aanmerkelijk onnauwkeuriger dan de verwachting met gebruikmaking van de eerste vijf scores van de analyse en $T_{x,-1}$.

3. Schatting van T_x met behulp van 500 mbar hoogtes.

In het voorafgaande hoofdstuk zijn regressievergelijkingen besproken waarbij een aantal scores als ingangsparameters optraden. Het is echter ook mogelijk om dezelfde Perfect Prog methode rechtstreeks toe te passen op de 500 mbar hoogtes in een aantal roosterpunten eventueel gecombineerd met de ingangsparameter $T_{x,-1}$.

De regressievergelijkingen hebben dan de gedaante:

$$T_x = \overline{T}_x + \sum_i C_i . (h_i - \overline{h}_i)$$

respectievelijk

$$T_x = \overline{T}_x + C_o . (T_{x,-1} - \overline{T}) + \sum_j C_j . (h_j - \overline{h}_j)$$

waarin h_j = hoogte van het 500 mbar-vlak in het j^e roosterpunt.

$\overline{h}_j$ = gemiddelde hoogte van het 500 mbar-vlak in het j^e roosterpunt.

Dergelijke regressievergelijkingen zijn ook op het BK3-materiaal uit de maanden augustus en september 1974 toegepast voor het bepalen van T_x. De resultaten hiervan zijn gegeven in tabel 4.V.

Tabel 4.V: Resultaten van schatting T_x uit BK3-prognoses en -analyses (hoogtes).

		Regressievergelijkingen met $T_{x,-1}$		Regressievergelijkingen zonder $T_{x,-1}$	
	aantal gevallen	$\Sigma\|T_o - T_c\|$ analyse	$\Sigma\|T_o - T_c\|$ prog.	$\Sigma\|T_o - T_c\|$ analyse	$\Sigma\|T_o - T_c\|$ prog.
aug.	23	32	39	42	45
sept.	20	24	36	31	50
Totaal	43	57	75	73	95

Het blijkt, dat de temperatuurschattingen zonder gebruik van $T_{x,-1}$ aanmerkelijk onnauwkeuriger zijn dan de schattingen van de weerkamermeteorologen. Indien wel gebruik wordt gemaakt van $T_{x,-1}$ dan blijken de verwachtingen m.b.v. de hoogtes van de analyse even nauwkeurig als de verwachtingen op de scores van de analyse (tabel 4.III en 4.V). Bij het gebruik van BK3-prognoses blijken de temperatuurschattingen die berekend zijn op grond van de scores iets nauwkeuriger dan de temperatuurverwachtingen op grond van de hoogtes. Dit verschil wordt waarschijnlijk veroorzaakt door het feit dat de BK3 de score van de vijfde eigenvector, die belangrijk is voor de verwachting van de maximumtemperatuur, tamelijk nauwkeurig kan voorspellen. De correlatie tussen opgetreden en voorspelde score van de vijfde eigenvector is 0.80; bij de eerste eigenvector is deze correlatiecoëfficiënt slechts 0.65. Waarschijnlijk wordt door het gebruik van de scores op zinvollere wijze van de BK3-prognose gebruik gemaakt dan wanneer de voorspelling van T_x rechtstreeks via de hoogtes van het 500 mbar-vlak verloopt. Dit hangt vermoedelijk samen met het feit dat door de BK3 patronen beter worden voorspeld dan absolute hoogtes. Belangrijk is om op te merken dat dit een BK3-effect is (op de analyses doen beide methodes het bijna evengoed); een ander numeriek model kan dus een ander resultaat opleveren.

4.3 Selectie analoge stromingspatronen.

1. Inleiding.

Bij het opstellen van de "meerdaagse verwachting" (verwachting van twee en drie dagen vooruit) wordt gebruik gemaakt van analoge stromingssituaties die zich in het verleden hebben voorgedaan. Het stromingspatroon van de uitgangsdag (analyse) en het stromingspatroon twee en drie dagen vooruit (prognoses van het Amerikaanse zeslagen model) zijn bekend. Uit een set grondkaarten en 500 mbar-kaarten van 1949 tot heden worden door de meteoroloog situaties geselecteerd die zoveel mogelijk overeenkomen met de prognoses.

Er is een begin gemaakt om de visuele analogenselectie door een automatische selectie m.b.v. de computer te vervangen. Daartoe zijn een aantal selectiemethoden uitgeprobeerd, die bij een gegeven uitgangspatroon een analoog patroon uit het verleden opzoeken. Bij deze selectiemethoden wordt alleen de momentane analogie beoordeeld.

De analogen zijn geselecteerd uit een bestand van 500 mbar-gegevens van de jaren 1949 t/m 1970; op een gedeelte van dit bestand (1961 t/m 1970) zijn de E.O.F. berekend zoals beschreven is in hoofdstuk 3.

Voor een viertal prognoses zijn volgens een aantal methodes analogen geselecteerd.

De geselecteerde analogen zijn daarna nog door een meteoroloog beoordeeld. De beoordeling vond plaats in drie klassen; Goed (+), Matig (0) of Slecht (-). In tabel 4.VI* zijn de analogen gegeven voor zover de beoordeling + of +/0 was.

De analogen zijn geselecteerd uit dagen waarvan het dagnummer in het jaar minder dan twintig verschilt met het dagnummer van de prognose. Dit komt overeen met de verschuiving die ook door de meteorologen wordt toegestaan.

* (Tabel 4.VI, zie blz. 31).

2. Beschrijving methodes analogen selectie.

1. Van het 500 mbar-patroon van de prognose zijn de eerste p scores berekend, nl. $\alpha_1, \alpha_2, \ldots\ldots, \alpha_p$.

Van de dagen waaruit de analoge situatie moet worden geselecteerd worden ook de scores berekend. Stel de scores van de n^e dag uit het bestand zijn: $\beta_{in}, \beta_{2n}, \ldots\ldots, \beta_{pn}$. De dag waarvoor de grootheid

$$D_n = \sum_{i=1}^{p} g_i \cdot (\alpha_i - \beta_{in})^2$$

minimaal is, wordt als de beste analoog beschouwd. De dag met de één na kleinste waarde van D_n wordt als de één na beste analoog beschouwd, etc. De twintig dagen met de beste analogie worden in de computeruitvoer gegeven.

In het gedeelte van het bestand van 1949 t/m 1958 blijken in het "grote veld" een aantal punten systematisch te ontbreken terwijl het gebied van het kleine veld (zie fig. 3.1) over het gehele bestand compleet is. Om deze reden is de analoge selectie voor zover E.O.F. worden gebruikt op het kleine veld uitgevoerd.

De selectie is twee maal gedaan:

A. Met de gewichten $g_1=g_2= \ldots =g_{20}=1$ en $g_{21}=g_{22} \ldots =g_{30} = 0$. Analogen selectie die op deze wijze is uitgevoerd zal worden aangeduid met berekening I. De bruikbare analogen die van de vier prognoses zijn gevonden staan in kolom 4 tabel 4.VI.

B. Met de gewichten $g_1=g_3=5$; $g_2=g_4=g_5= \ldots =g_{20}=1$; g_{21} t/m $g_{30} = 0$.

Het blijkt dat voor het kleine veld de eerste en derde eigenvector een groot gedeelte van de variantie van het 500 mbar-vlak in de omgeving van Nederland opleveren.

Daarom is de selectie nogmaals uitgevoerd maar met grotere gewichten voor de eerste en derde score. Deze methode van analogen selectie wordt aangegeven met berekening II. De resultaten staan vermeld in tabel 4.VI, kolom 5.

Tabel 4.VI: Resultaten analogen selektie.

1	2	3	4	5	6	7	8	9	10
72 uur prog, 00.00h voor:	analogen	beoordeling meteoroloog	I	II	III	IV	V	VI	VII
26-10-74 00.00 z	6-11-52	0/+	11	-	11	11	5	10	8
	9-11-58	+	8	-	8	4	6	8	6
	17-10-61	+	(+1)16	-	(+1) 7	15	20	3	4
	5-10-63	0/+	19	(-1)11	(+1) 4	-	-	-	-
31-10-74 00.00 z	27-10-50	0/+	-	-	-	15	17	11	10
	13-11-52	0/+	(-1)11	(-1)14	(-1)13	-	(-1)19	-	18
	8-11-60	0/+	2	(+1)13	2	1	1	1	2
	13-10-64	0/+	(+1) 9	10	(+1)12	(+1)10	(+1) 9	(+1) 6	(-1) 9
	16-10-64	0/+	(-1)17	-	(-1)11	(-1) 6	(-1) 6	(-1) 3	17
2-11-74 00.00 z	13-10-50	+	-	-	-	-	-	7	19
	19-11-54	+	2	3	3	19	7	14	10
	7-11-56	+	17	-	20	-	-	-	-
	18-10-60	+	1	1	2	20	5	17	9
3-11-74 00.00 z	31-10-49	0/+	1	7	-	2	2	6	1
	1-11-49	0/+	2	6	-	4	1	8	2
	18-10-60	+	6	19	12	13	8	(+1)20	16
	12-11-68	0/+	(-1) 8	-	(+1)16	14	(-1)11	5	7
	Aantal gemiste gevallen		2	8	4	4	3	3	2
	Aantal gevallen met één dag verschuiving		5	3	6	3	4	3	1

Kolom 1 : Datum van de onderzochte situaties.

Kolom 2 : Data van de bruikbare analogen.

Kolom 3 : Beoordeling van de analogen door de meteoroloog.

Kolom 4 t/m 10 : Rangnummer gegeven door aangeduide methode aan de onder 2 genoemde dag.

- Rangnummer groter dan 20.

(-1) Rangnummer van de dag voorafgaand aan die genoemd onder 2.

(+1) Rangnummer van de dag volgend op die genoemd onder 2.

Aanduiding (-1) of (+1) houdt tevens in dat dag zelf geen rangnummer kleiner dan 20 heeft.

2. In het voorafgaande is er naar gestreefd om analogen te selecteren door de afstand tussen de score vector van de uitgangstoestand en de score vector van de analoge dag minimaal te maken. Een andere methode om analogen te zoeken is dat de score vector van de prognose zoveel mogelijk parallel is aan de score vector van de analoge dag, d.w.z. het inproduct tussen de score vectoren moet maximaal zijn. Dit komt overeen met de eis dat

$$I_n = \frac{\sum_{i=1}^{p} \alpha_i \beta_{in}}{\sum_{i=1}^{p} \alpha_i^2 \cdot \sum_{i=1}^{p} \beta_{in}^2}$$

maximaal moet zijn.
Meteorologisch betekent dit, dat vooral op de posities van de druksystemen wordt gelet en minder op de intensiteit van de systemen. De bruikbare analogen zijn in kolom 6, tabel 4.VI gegeven. Deze selectiemethode zal worden aangeduid met berekening III.

3. Behalve het selecteren van analogen m.b.v. scores kan men voor de selectie ook rechtstreeks uitgaan van de hoogtes in de roosterpunten. Stel h_{oj} is de hoogte van het 500 mbar-vlak in het j^e roosterpunt van de uitgangstoestand en h_{nj} is de hoogte van het 500 mbar-vlak in het j^e roosterpunt op dag n, dan is

$$D_n = \sum_{j=1}^{63} (h_{oj} - h_{nj})^2$$

een maat voor de "afstand" tussen de velden.
De dag waarvoor D_n minimaal is, levert de beste analoge situatie op.

Deze methode heeft als voordeel boven de selectie m.b.v. de scores dat het op eenvoudige wijze mogelijk is om, door het toevoegen van gewichtsfactoren, de analogie in bepaalde delen van het kaartgebied zwaarder te laten wegen dan in andere gedeelten. De "afstand" D_n wordt dan:

$$D_n = \sum_{j=1}^{63} g_j \cdot (h_{oj} - h_{nj})^2 .$$

Voor de roosterpunten waarvan gegevens uit het tijdvak 1949 t/m 1958 ontbreken is de gewichtsfactor 0 genomen.
De selectiemethode waarbij gebruik wordt gemaakt van de hoogtes is met twee sets gewichten uitgevoerd.

a. de gewichten in alle roosterpunten waarvan de hoogtes van de hele periode 1949 t/m 1970 aanwezig zijn 1 genomen (berekening IV; tabel 4.VI, kolom 7).
b. de gewichten in dertig roosterpunten in de omgeving van Nederland zijn drie genomen; buiten dit gebied zijn de gewichten hetzelfde als a. (berekening V; tabel 4.VI, kolom 8.)

4. Bij het selecteren van analogen zijn de meteorologen vooral geïnteresseerd in de vorm van het stromingspatroon en ze zijn in mindere mate geïnteresseerd in de absolute hoogte van de 500 mbar-isohypsen. Het is dan ook mogelijk dat voor een goede analoog de gemiddelde waarde van de 500 mbar-hoogtes in de 63 roosterpunten enige decameters afwijkt van de gemiddelde hoogte van de roosterpunten in de prognose. Om deze reden zijn ook analogen bepaald waarbij de standaardafwijking tussen de prognose en de analoog is berekend volgens de volgende formule.

$$D_n = \sum_j g_j \cdot (h_{nj} - h_{oj})^2 \quad - \frac{\sum_j g_j \cdot (h_{nj} - h_{oj})^2}{\sum_j g_j}$$

De analogenselectie is twee maal uitgevoerd.

A. alle gewichten g_i=1 indien de gegevens in het roosterpunt i compleet in het bestand aanwezig zijn; indien niet compleet: g_i=0 (berekening VI, tabel 4.VI).

B. de gewichten in dertig roosterpunten (roosterpunten van het kleine veld) zijn g=3 genomen. De overige gewichten als onder A.

De gevonden analogen staan vermeld in tabel 4.VI onder berekening VII.

4.4 Resultaten analogenselectie.

In tabel 4.VI wordt een overzicht gegeven in hoeverre bruikbare analogen bij een viertal prognoses m.b.v. de verschillende berekeningen zijn gevonden. Noch bij de computerselectie noch bij visuele naselectie zijn nog andere bruikbare analogen gevonden dan degene die in de tabel staan vermeld.

In een aantal gevallen verschillen de machinaal geselecteerde analogen één dag t.o.v. de visueel geselecteerde analogen. Dit is, gezien de grote persistentie van het 500 mbar-vlak niet erg verwonderlijk. Bovendien is dit in de praktijk niet storend omdat de meteoroloog toch de dagen rondom de analoog automatisch zal bekijken.

Hoewel een viertal uitgangstoestanden betrekkelijk gering is om de resultaten van de verschillende berekeningen met elkaar te vergelijken blijkt dat,

1) bij de berekening II veel bruikbare analogen worden gemist;

2) bij de resultaten van berekening I en VII slechts twee goede analogen worden gemist. Dat bij de berekening VII zo'n goede overeenstemming met de selectie van de meteorologen wordt verkregen duidt er wel op, dat de meteorologen de analogie vooral op het stromingspatroon in de omgeving van Nederland baseren.

3) in verschillende artikelen worden de E.O.F. bijzonder bruikbaar geacht voor het selecteren van analogen. Het blijkt echter dat het direct vergelijken van hoogtes in een aantal roosterpunten minstens zulke goede selectiemethodes kan opleveren als selectiemethodes m.b.v. E.O.F.

Literatuur

1. Anonymus : Climatic change. Report of a working group of the Commission for Climatology. Prepared by J.M. Mitchell a.o. Techn. note W.M.O. no. 79.

2. P. Bernadet : Représentation d'un champ au moyen des fonctions orthogonales naturelles. Ministère des Transports. NIT-Section III - Pièce no. 33.

3. P. Bernadet : Idem, no. 34.

4. W.J.A. Kuipers : An experiment on numerical classification of Scalar Fields. Időjárás 74 (1970) 296-306.

5. A.P. Dempster : Elements of continuous multivariate analysis. Addison-Wesley (1969).

6. D.N. Lawley and A.E. Maxwell : Factor analysis as a Statistical Method. Butterworths, London (1971) (2nd ed.).

7. M.G. Kendall and A. Stuart : The advanced theory of Statistics. Volume 3. Griffin, London (1968) (2nd. ed.).

8. C. Levert : Onderzoek naar de interdiurne variabiliteit van enkele meteorologische grootheden. Wet. Rapp. Kon. Ned. Meteor. Inst. W.R. 60-2.

9. W.H. Klein, F. Lewis and G.P. Casely : Automated nationwide forecasts of maximum and minimum temperature. Journ. Appl. Met. 6 (1967) 216.

10. H.C. Hamaker : On multiple regression analysis. Statistica Neerlandica 16 (1962) 31.

-o-o-o-

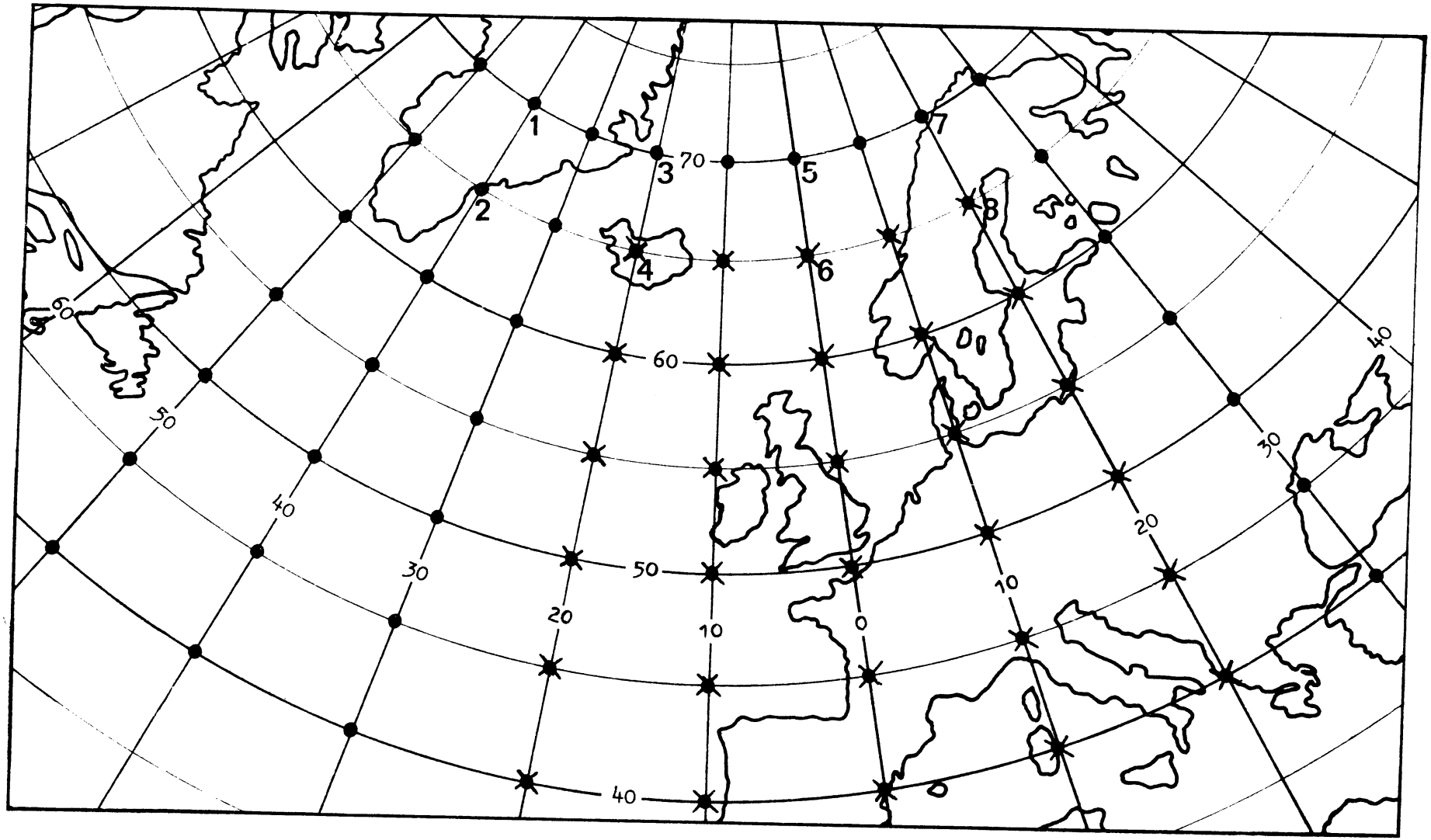

Fig. 3.1 Overzicht van de gebruikte roosterpunt configuraties.
De roosterpunten aangeduid met ● vormen het "grote veld".
De roosterpunten aangeduid met x vormen het "kleine veld".
Het veld van 55 roosterpunten ontstaat door uit het grote veld de roosterpunten 1 t/m 8 weg te laten.

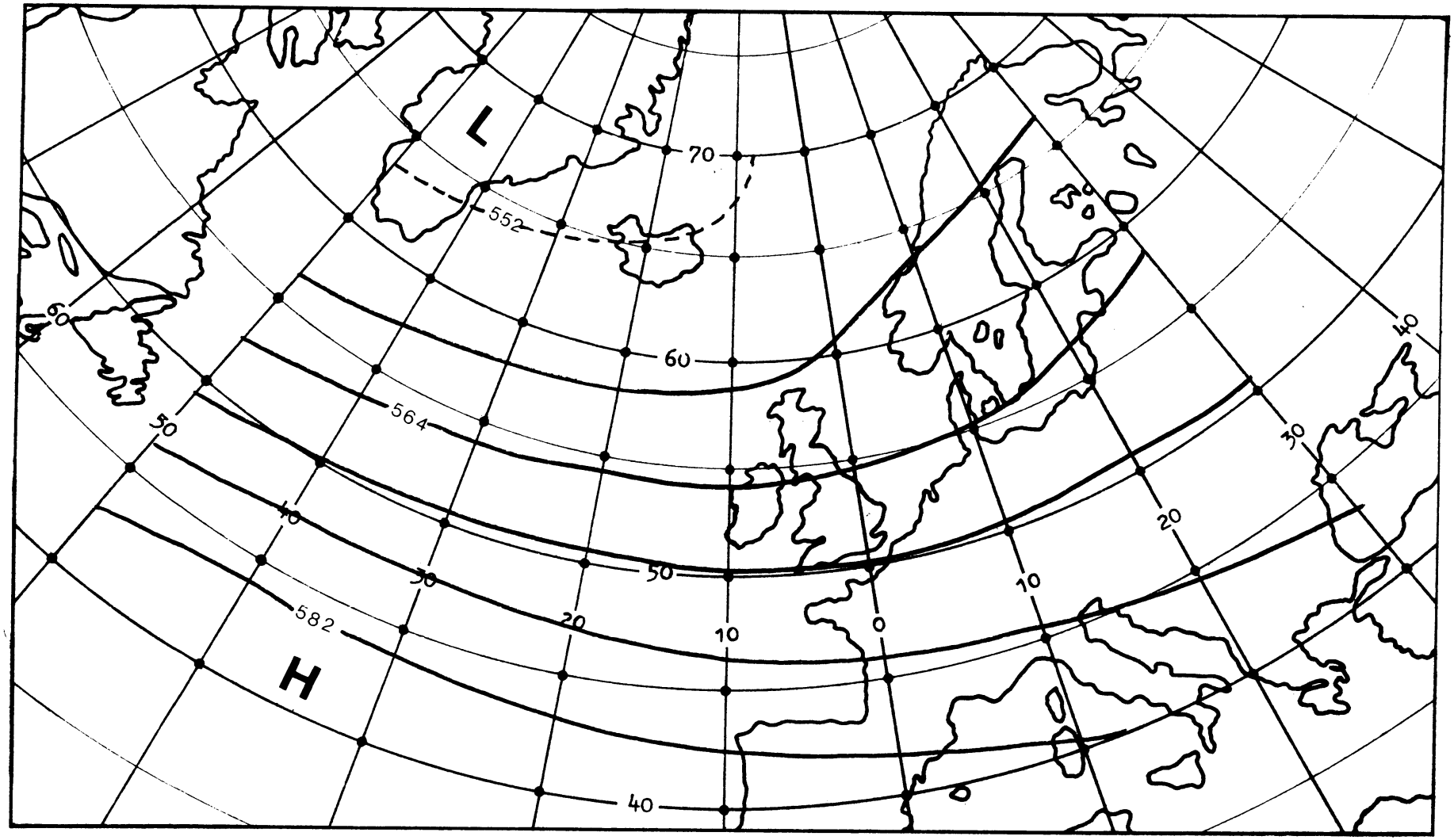

Fig. 3.2 Gemiddelde hoogteveld van augustus berekend uit de gegevens van de periode 1961 t/m 1970.

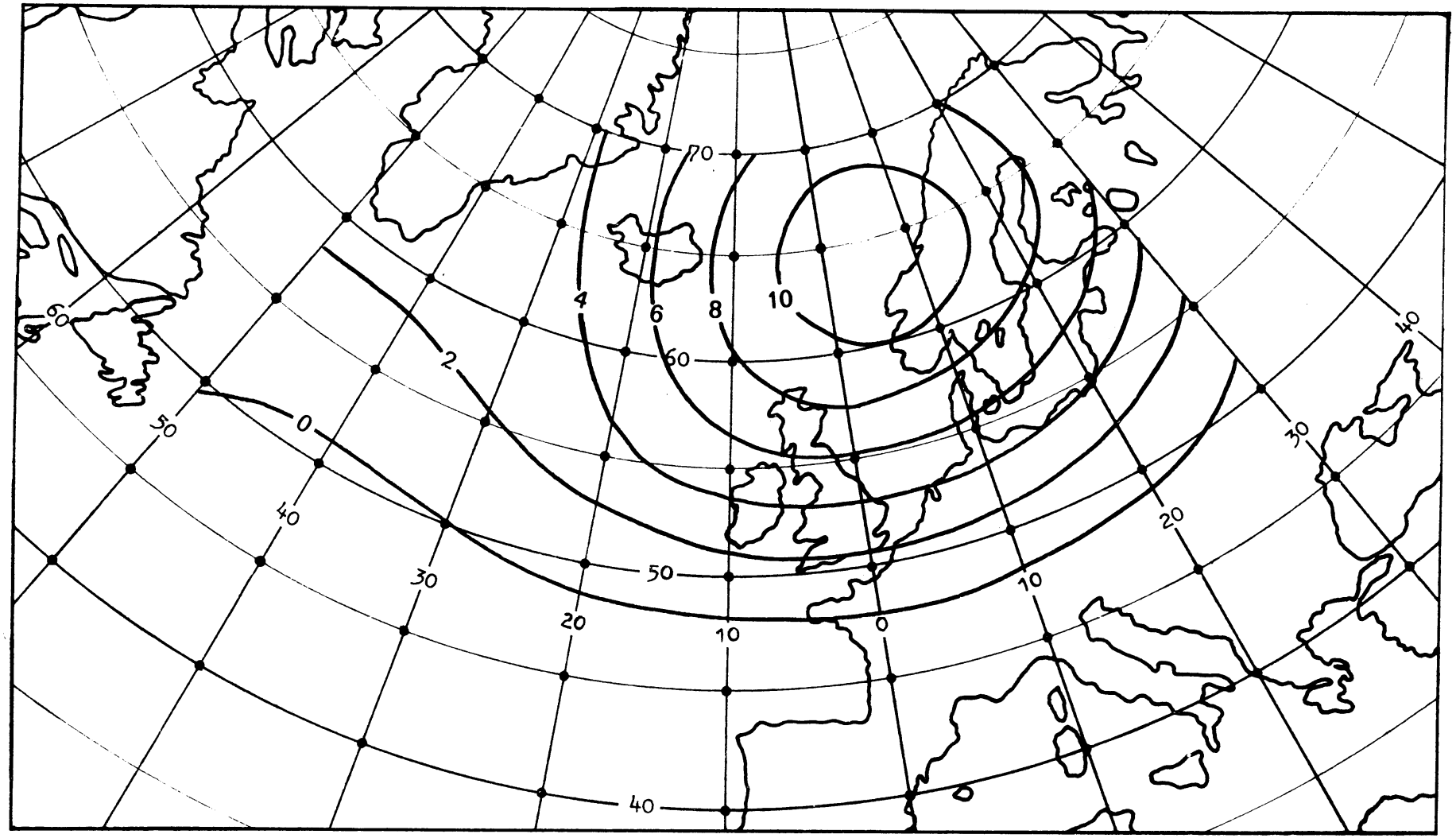

Fig. 3.3 le gewichtsvector voor augustus.

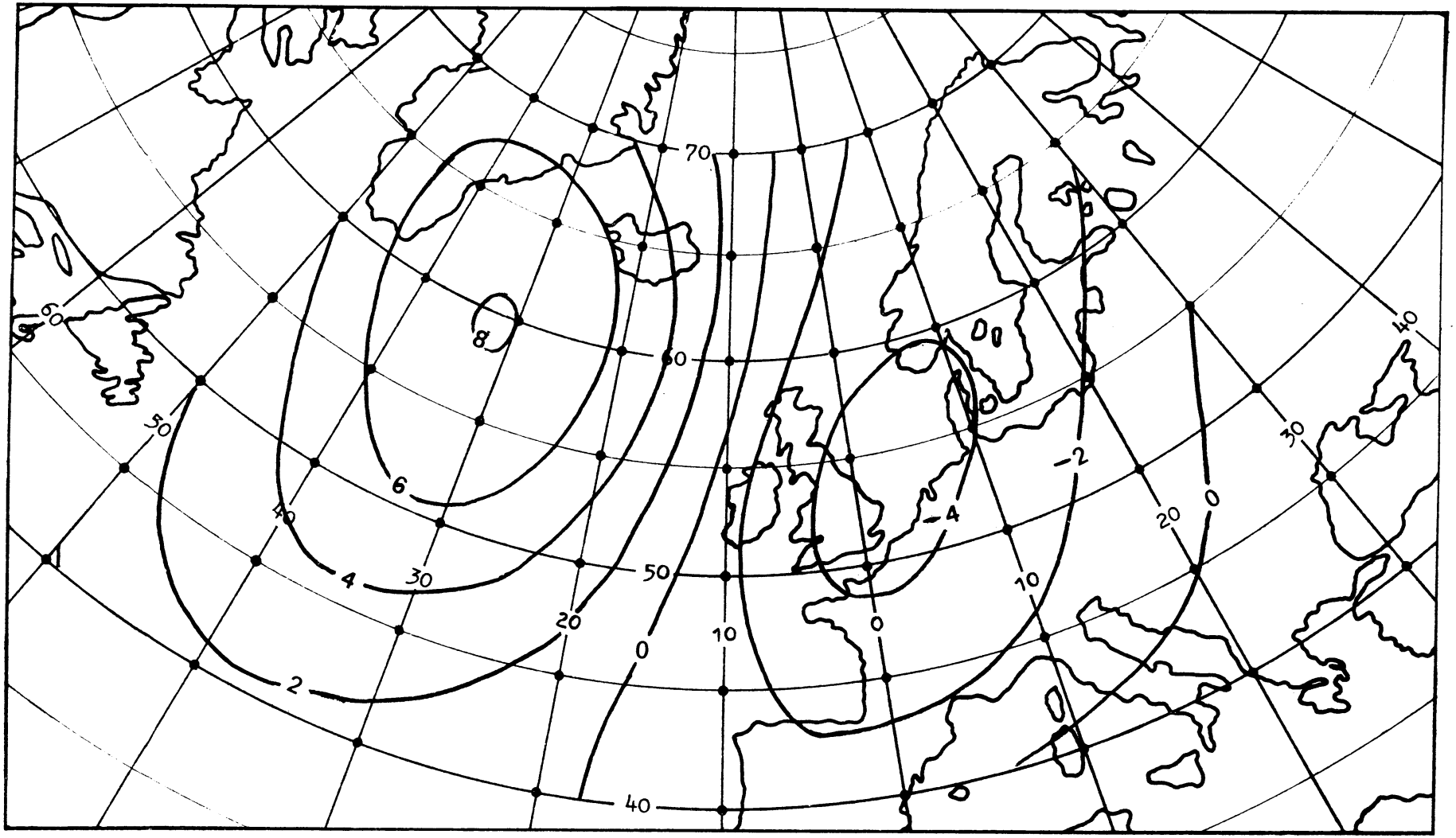

Fig. 3.4 2e gewichtsvector voor augustus.

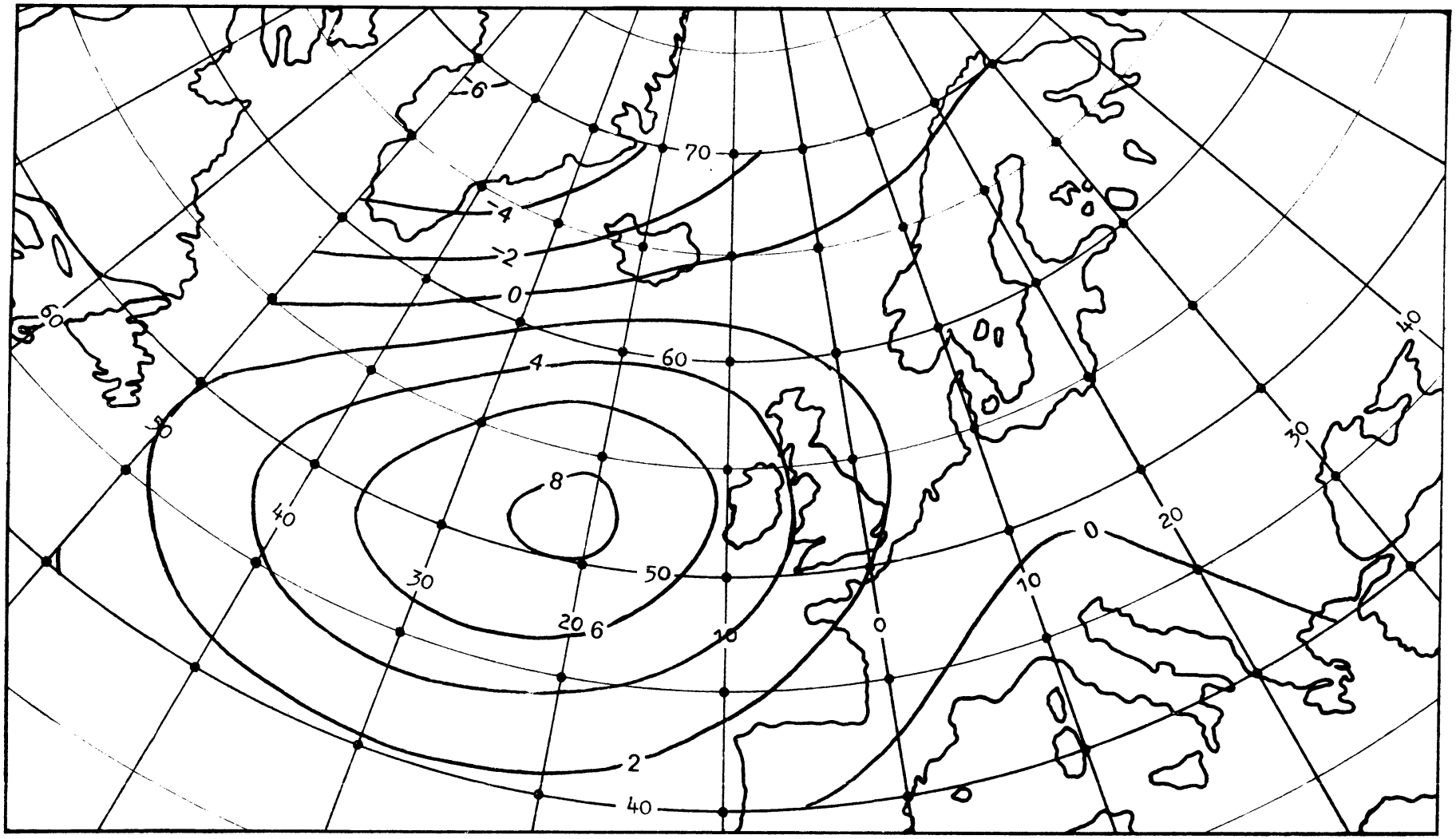

Fig. 3.5 3e gewichtsvector voor augustus.

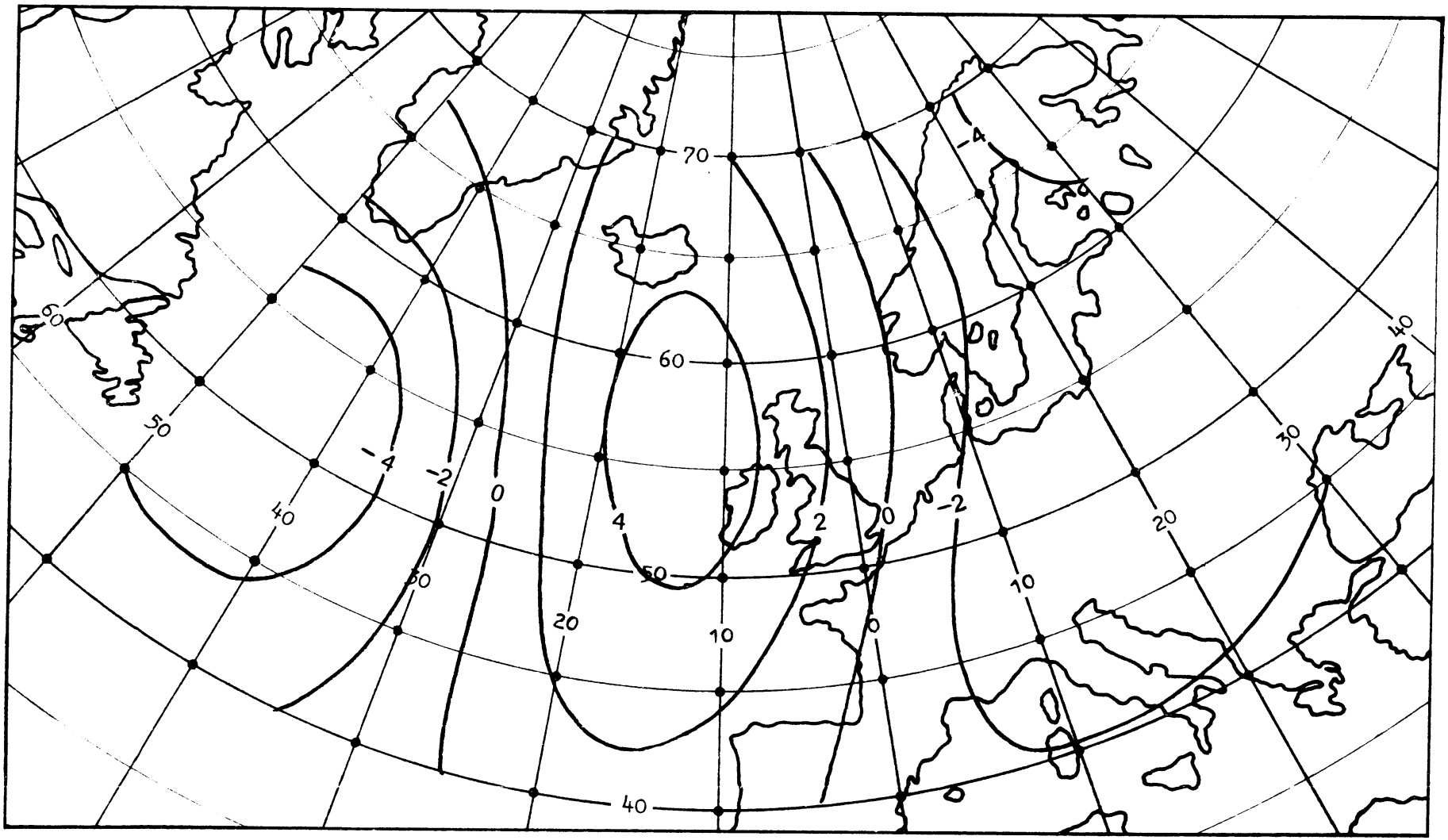

Fig. 3.6 4e gewichtsvector voor augustus.

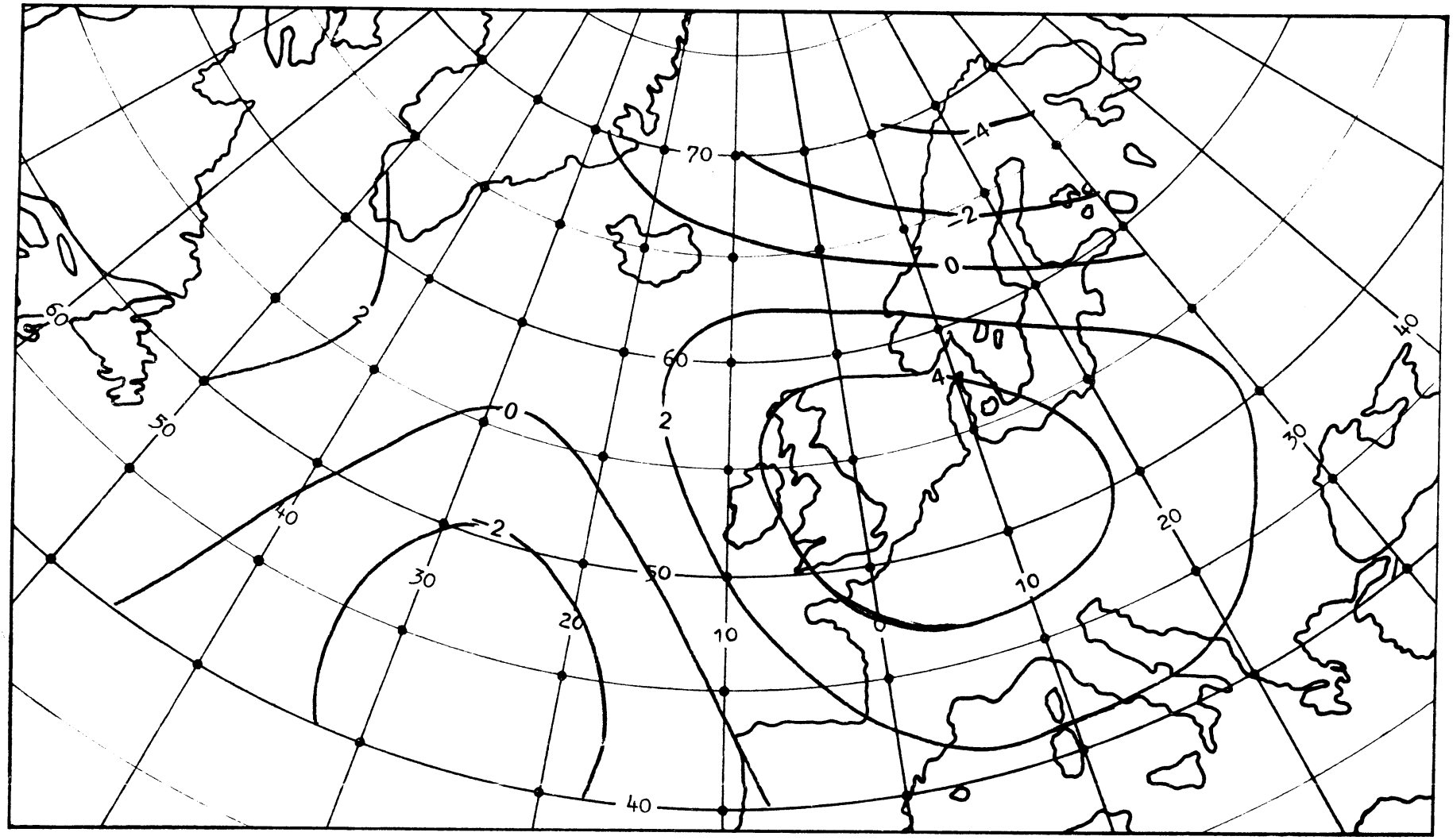

Fig. 3.7 5e gewichtsvector voor augustus.

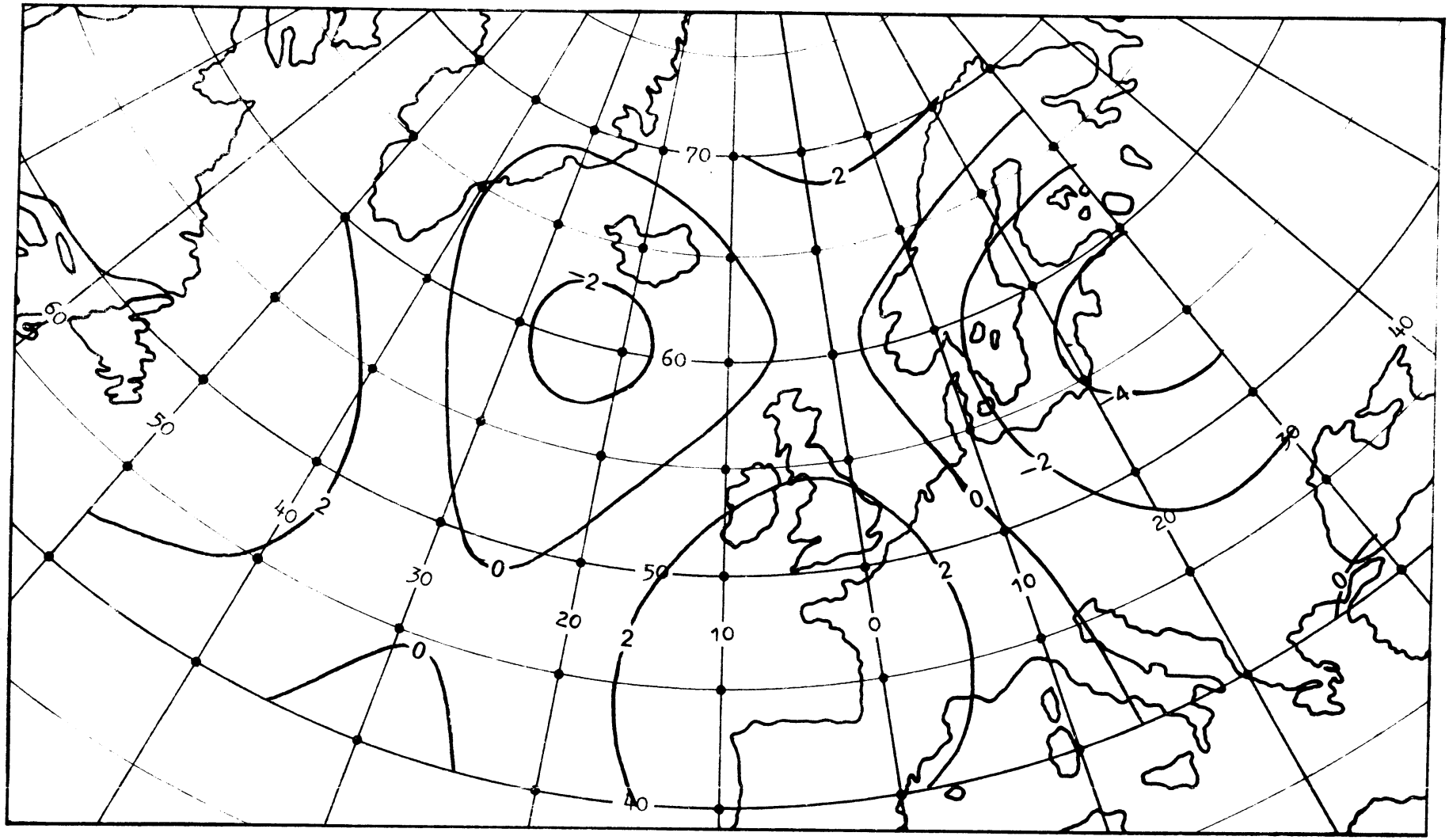

Fig. 3.8 6e gewichtsvector voor augustus.

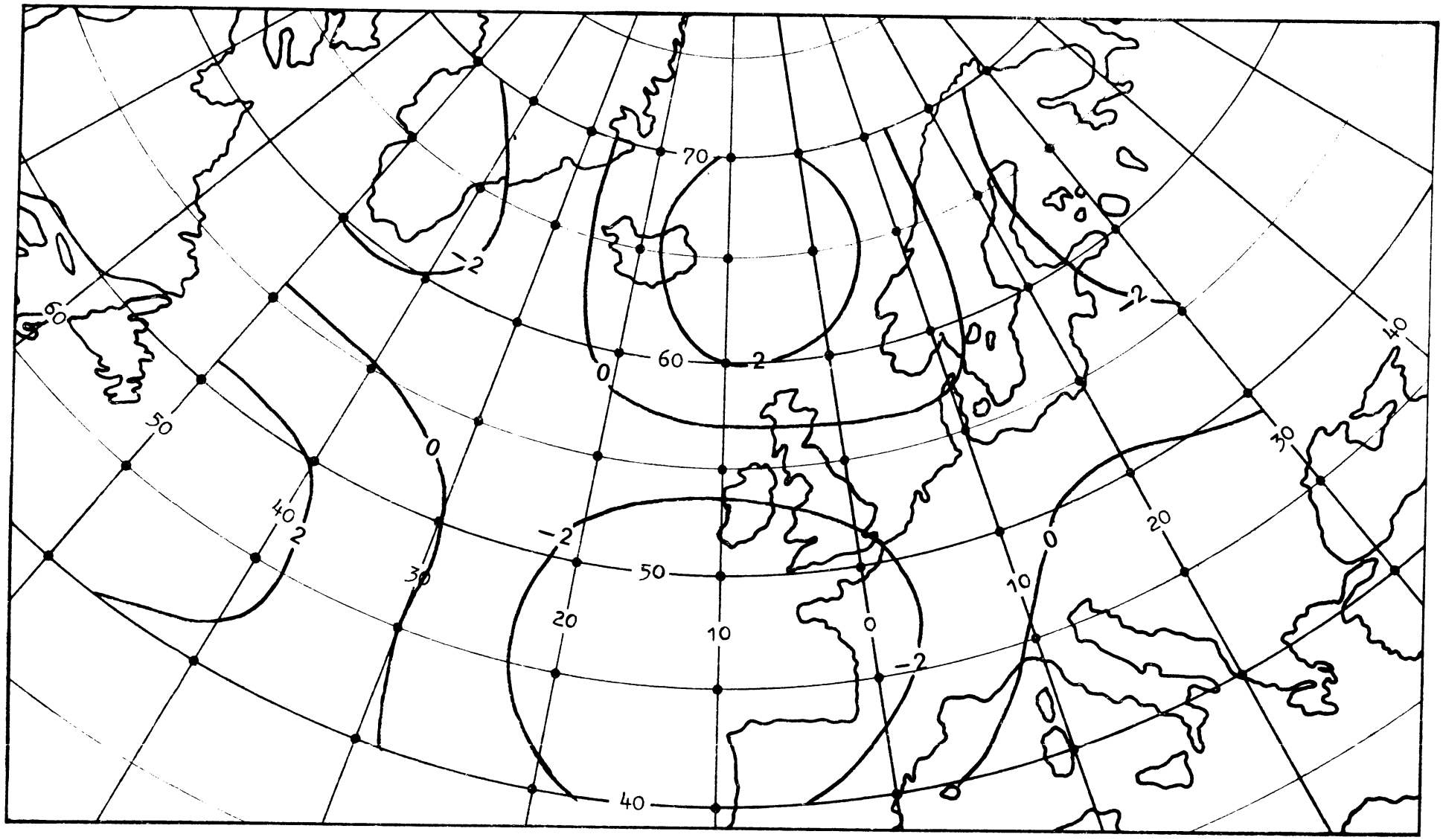

Fig. 3.9 7e gewichtsvector voor augustus.

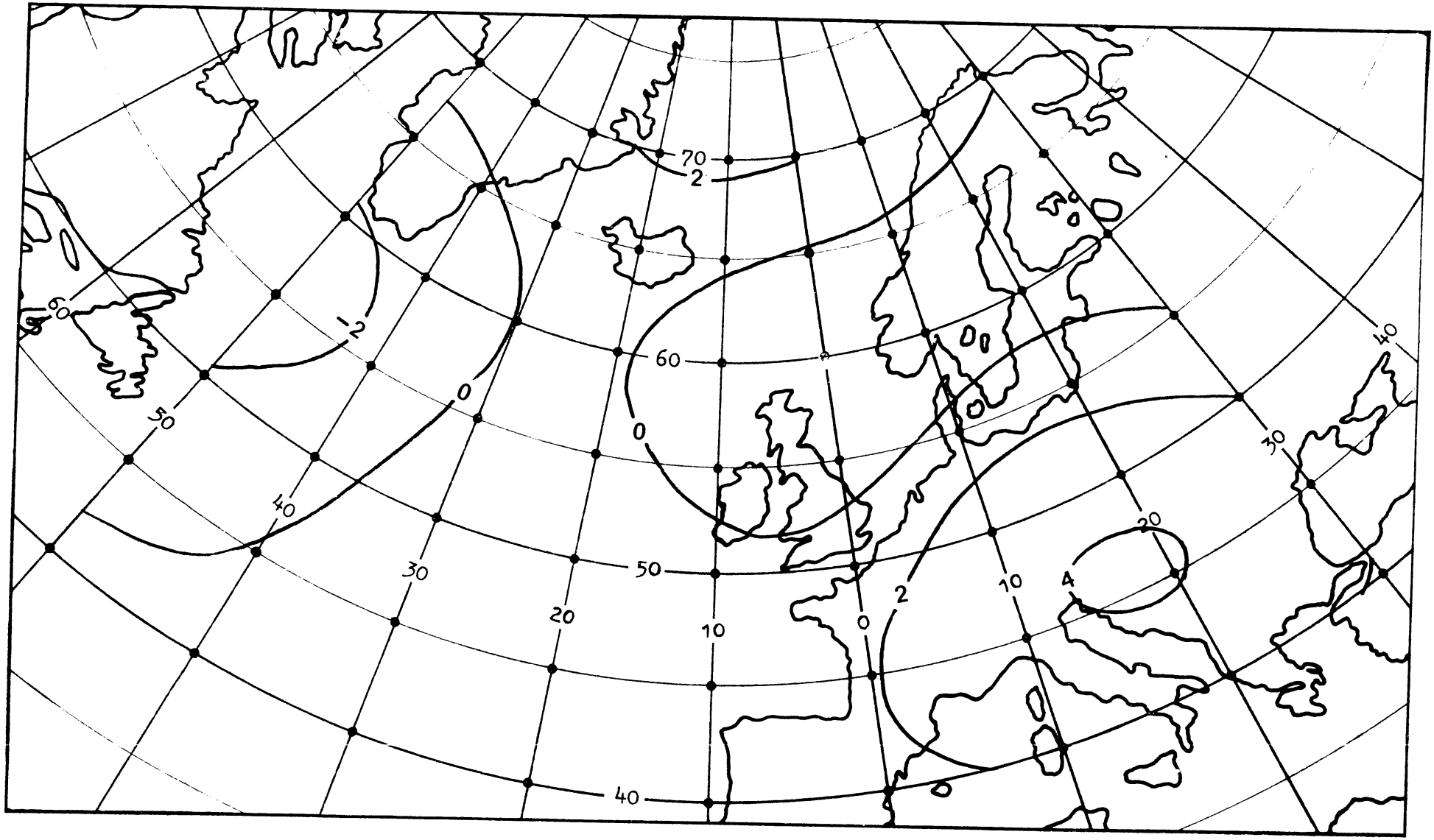

Fig. 3.10 8e gewichtsvector voor augustus.

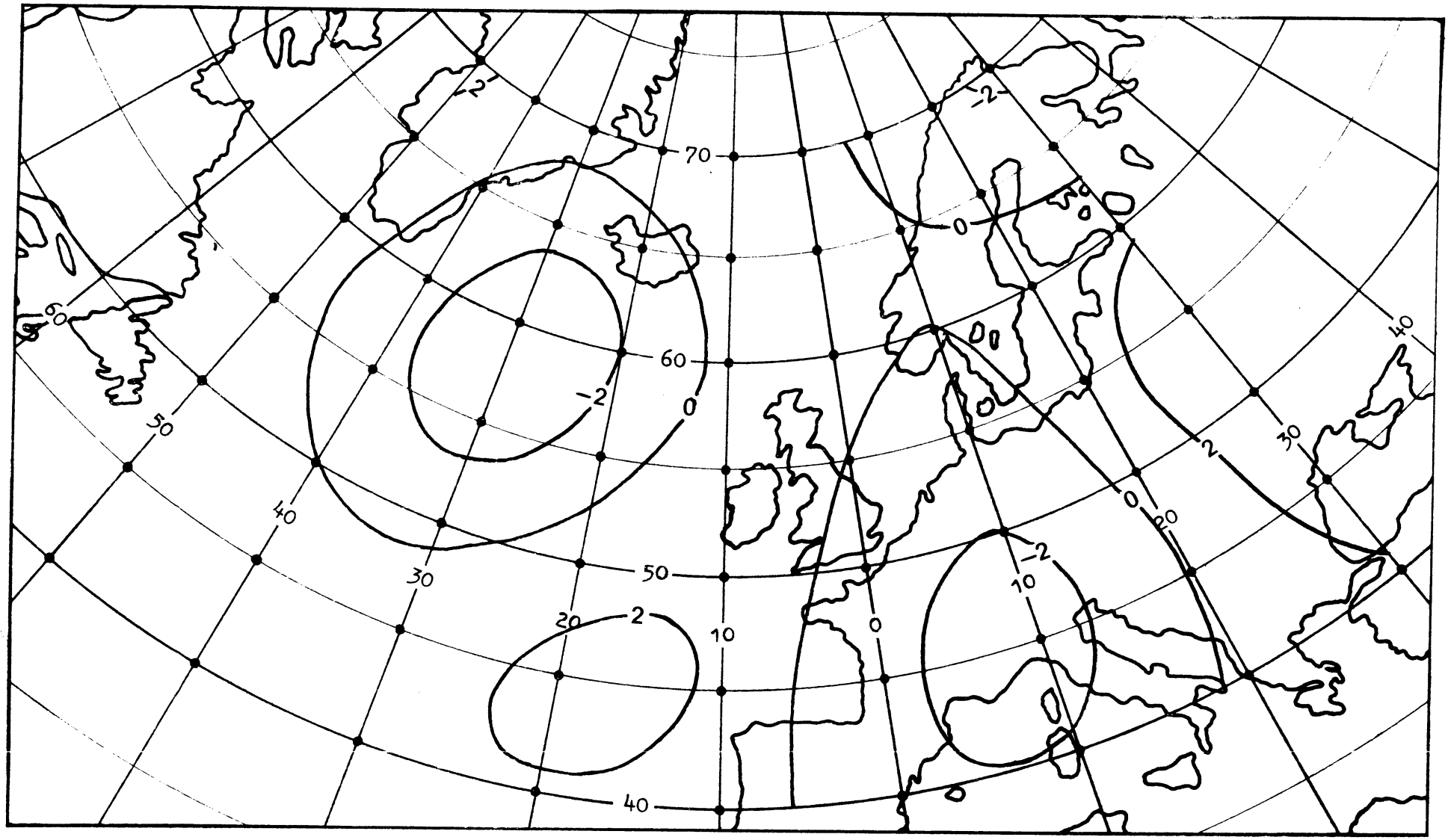

Fig. 3.11 9e gewichtsvector voor augustus.

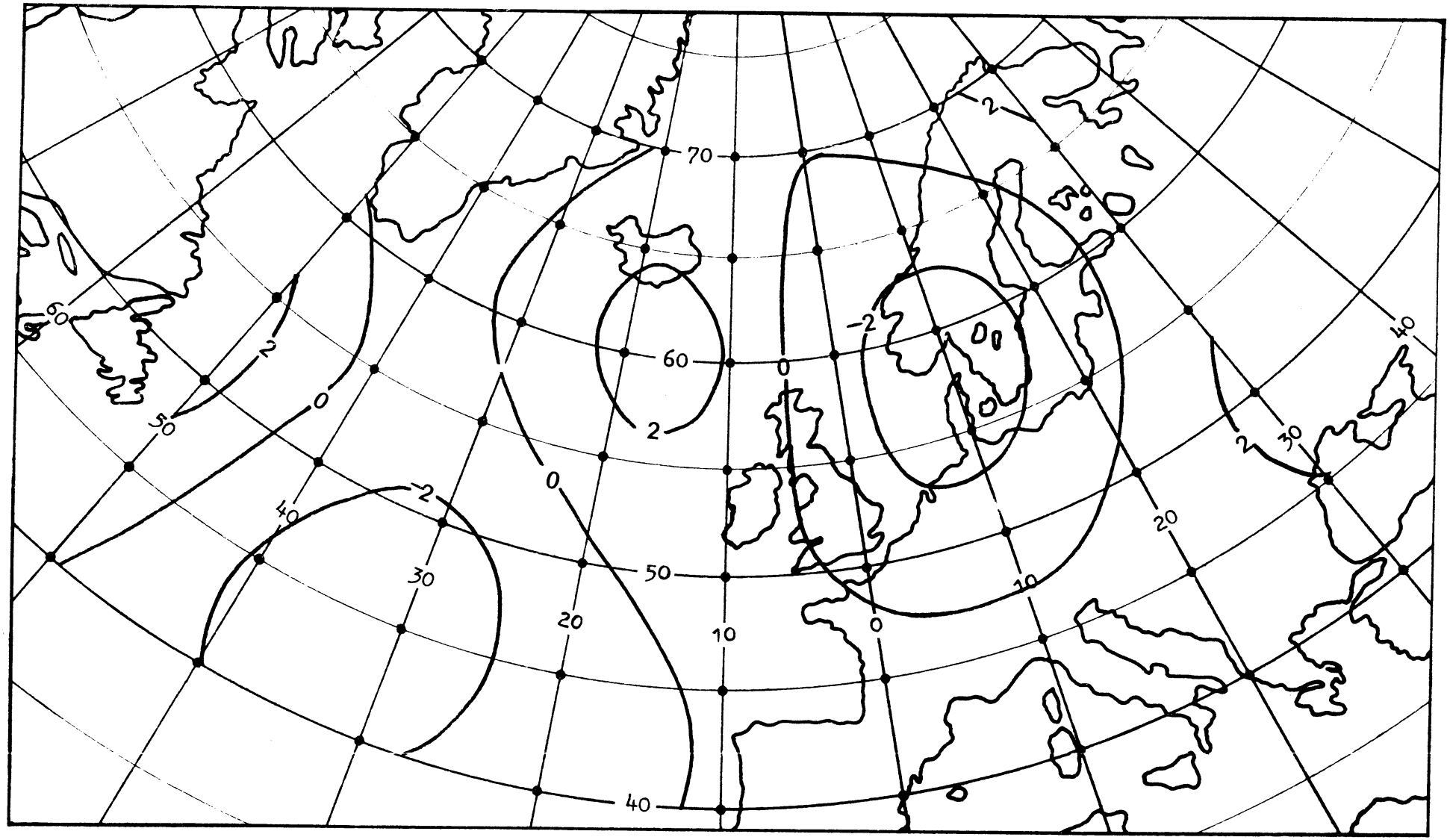

Fig. 3.12 10e gewichtsvector voor augustus.

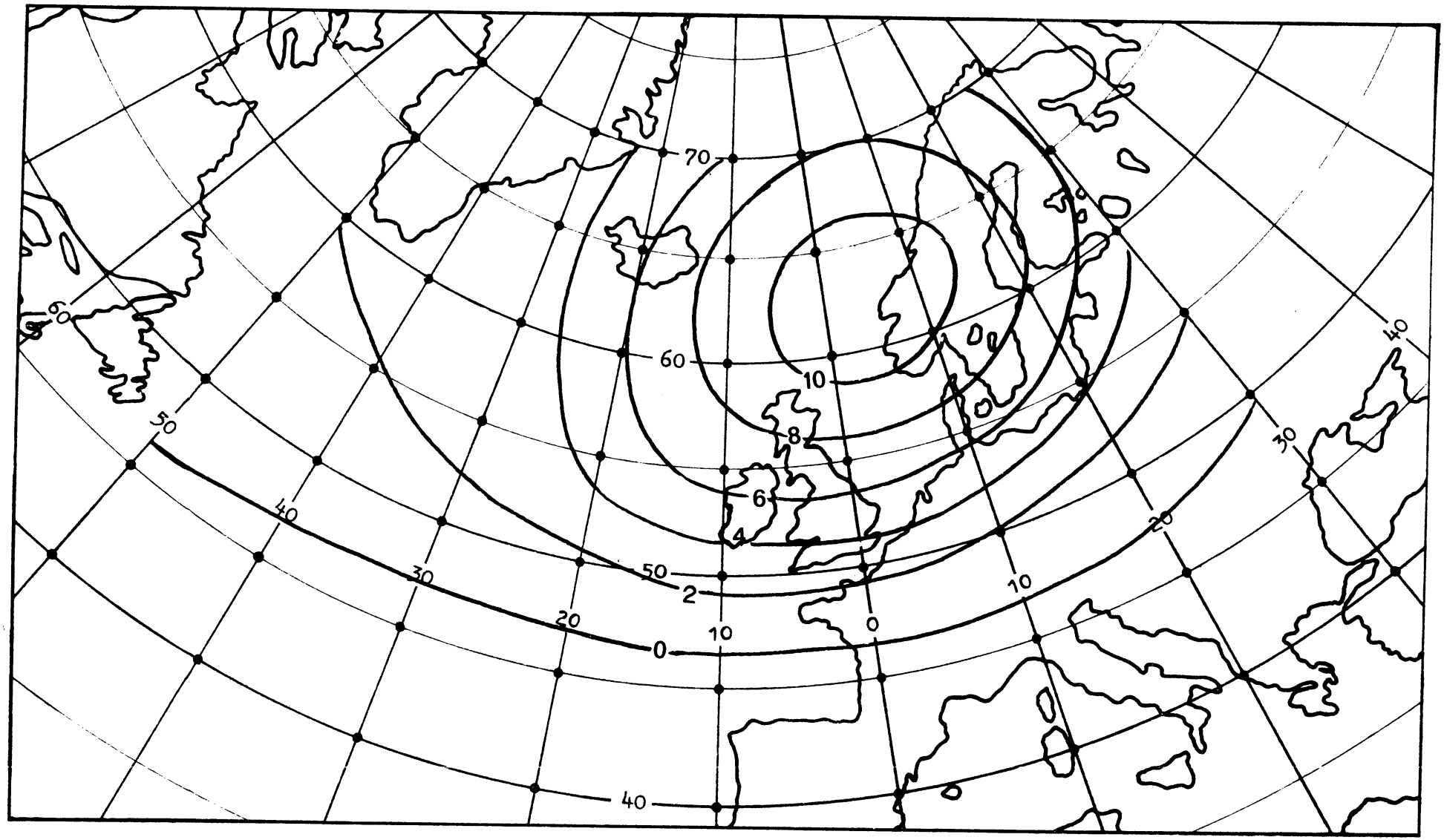

Fig. 3.13 1e gewichtsvector voor augustus op veld van 55 roosterpunten.

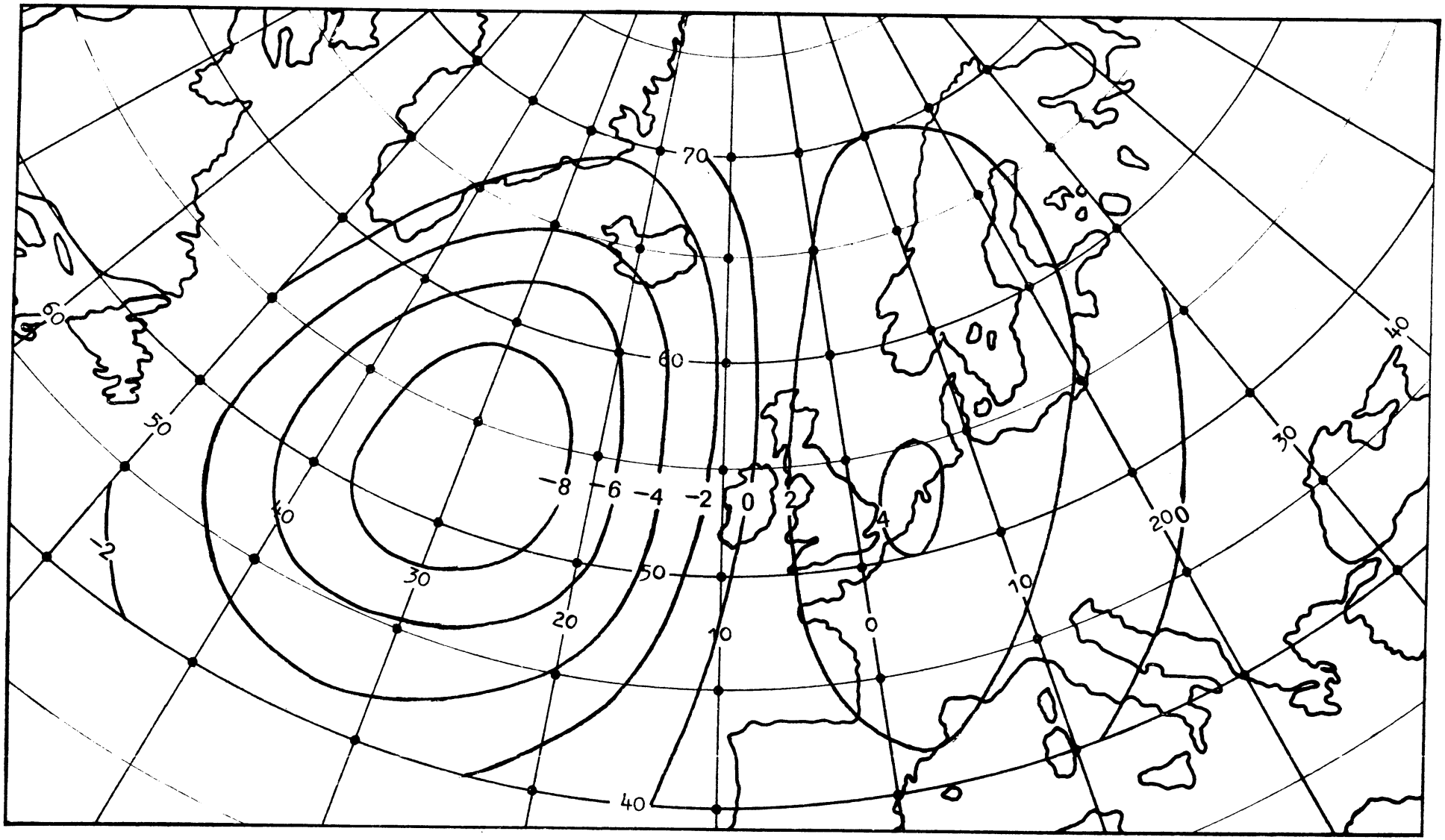

Fig. 3.14 2e gewichtsvector voor augustus op veld van 55 roosterpunten.

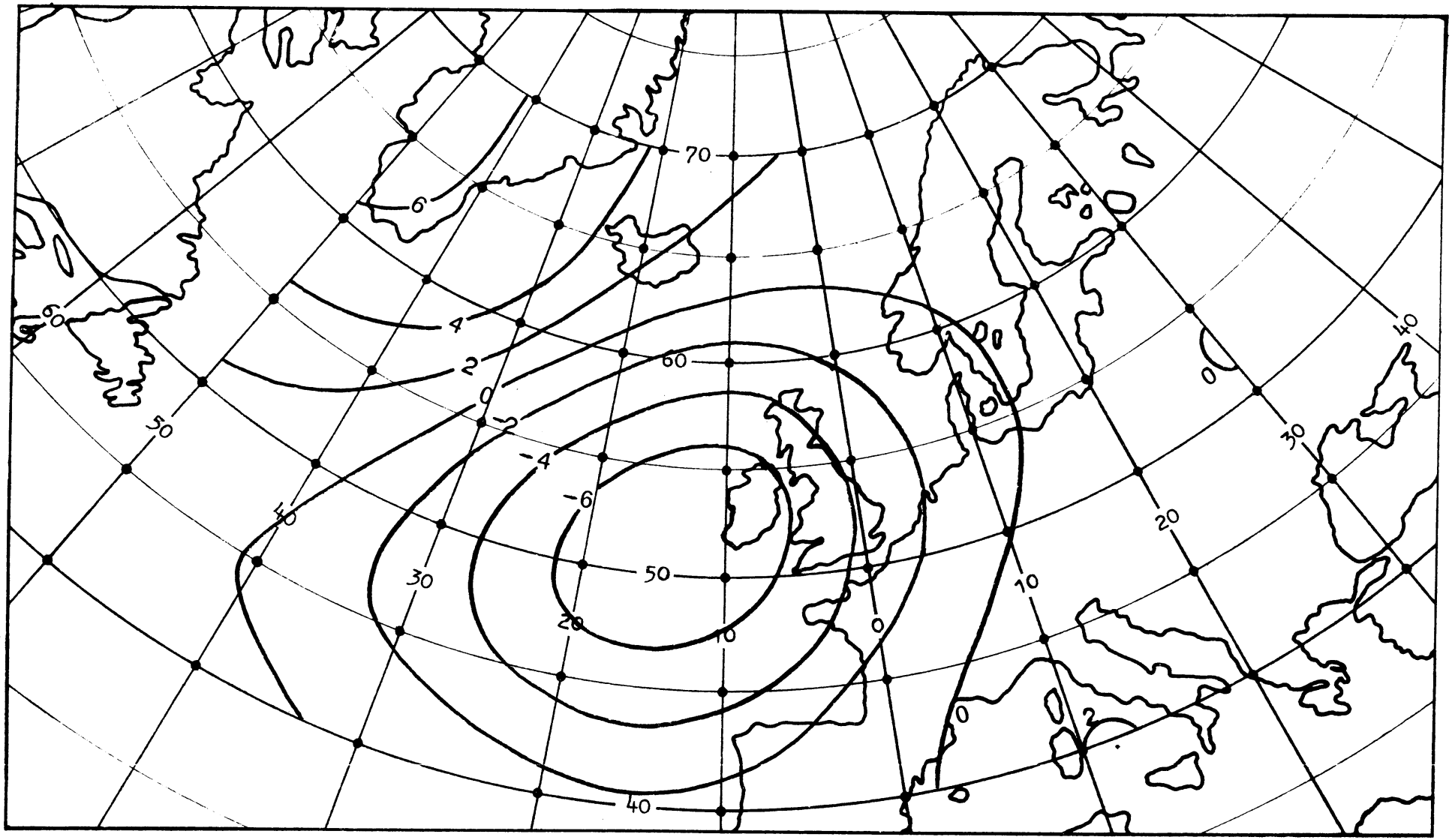

Fig. 3.15 3e gewichtsvector voor augustus op veld van 55 roosterpunten.

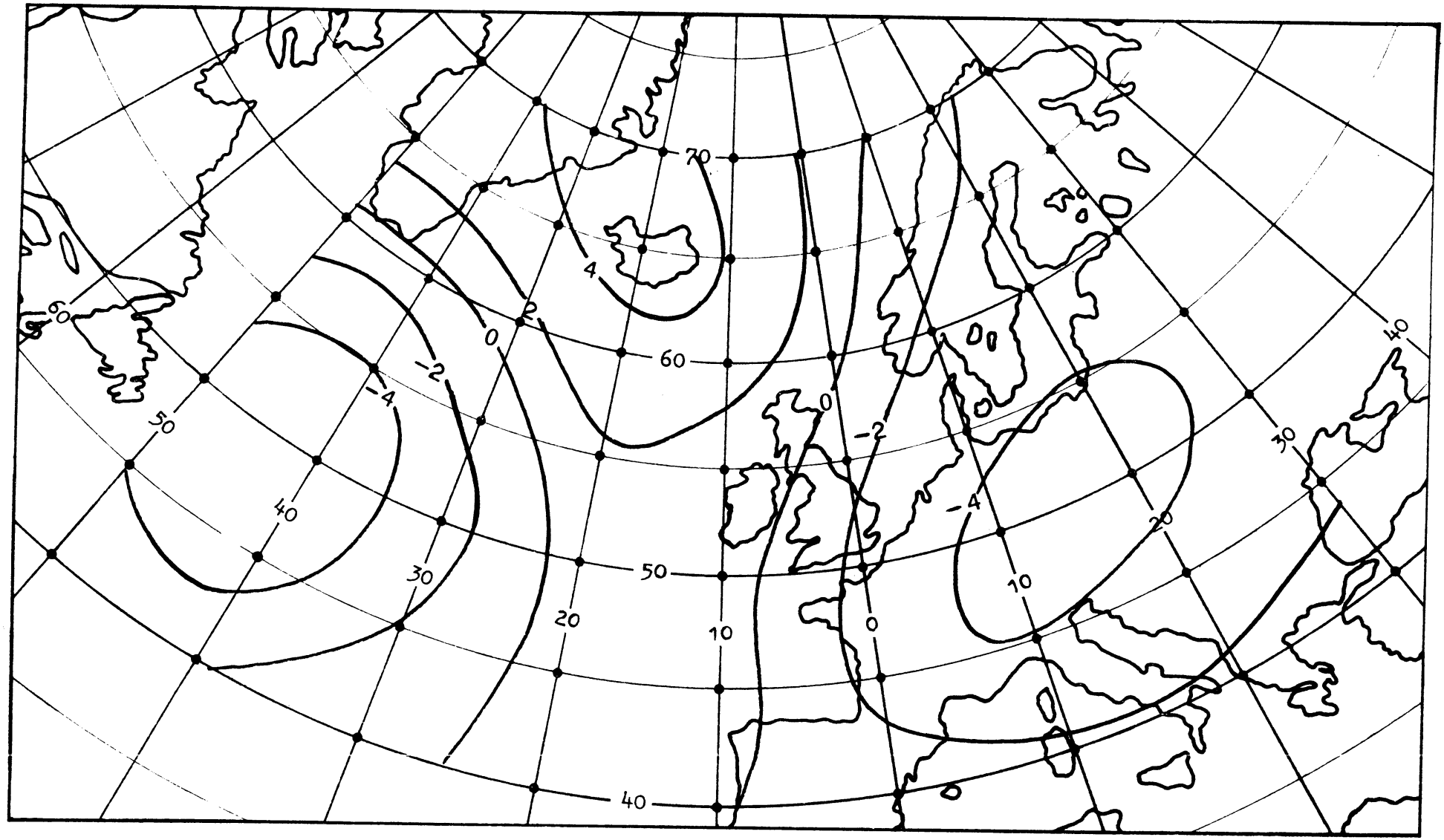

Fig. 3.16 4e gewichtsvector voor augustus op veld van 55 roosterpunten.

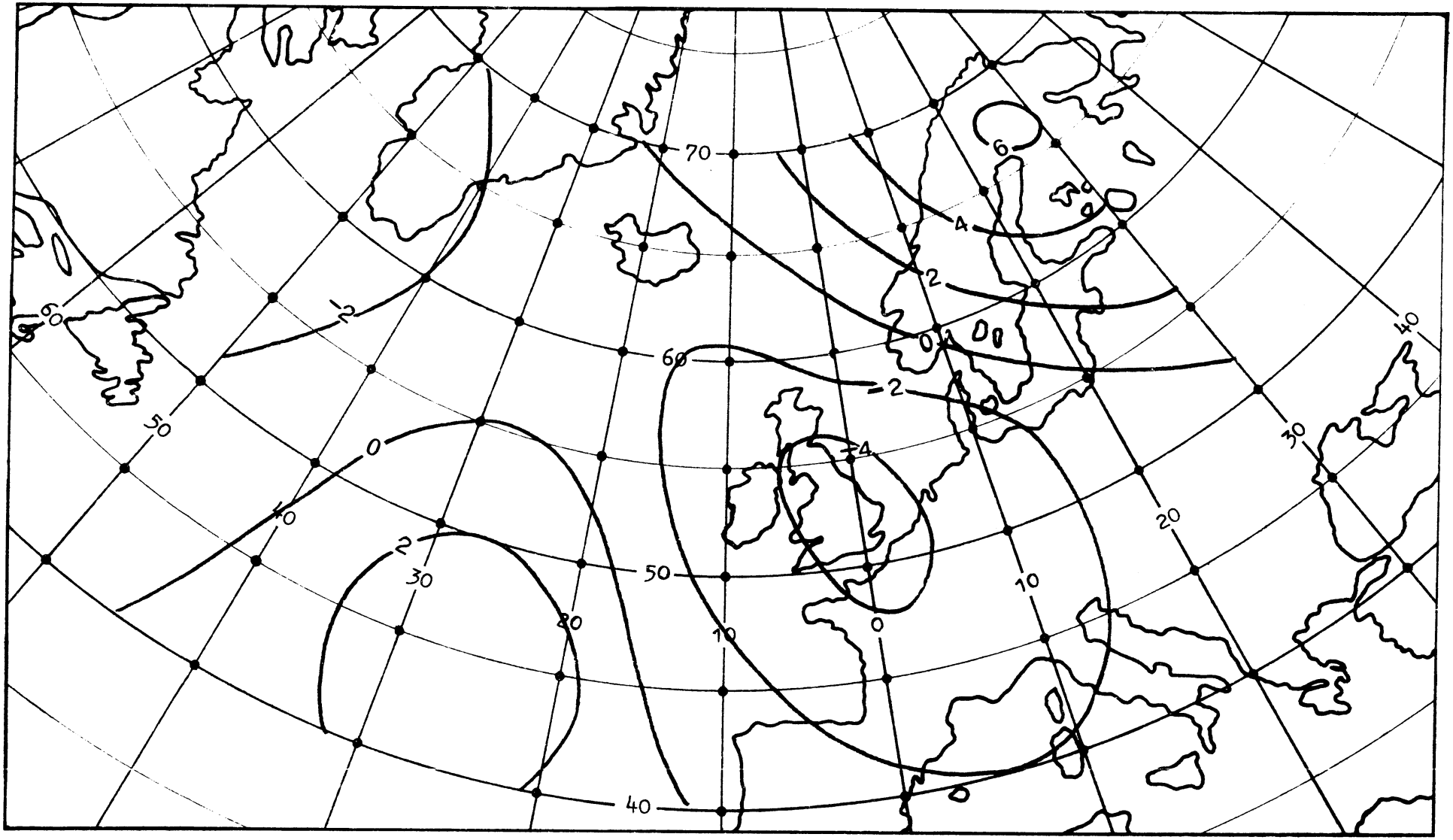

Fig. 3.17 5e gewichtsvector voor augustus op veld van 55 roosterpunten.

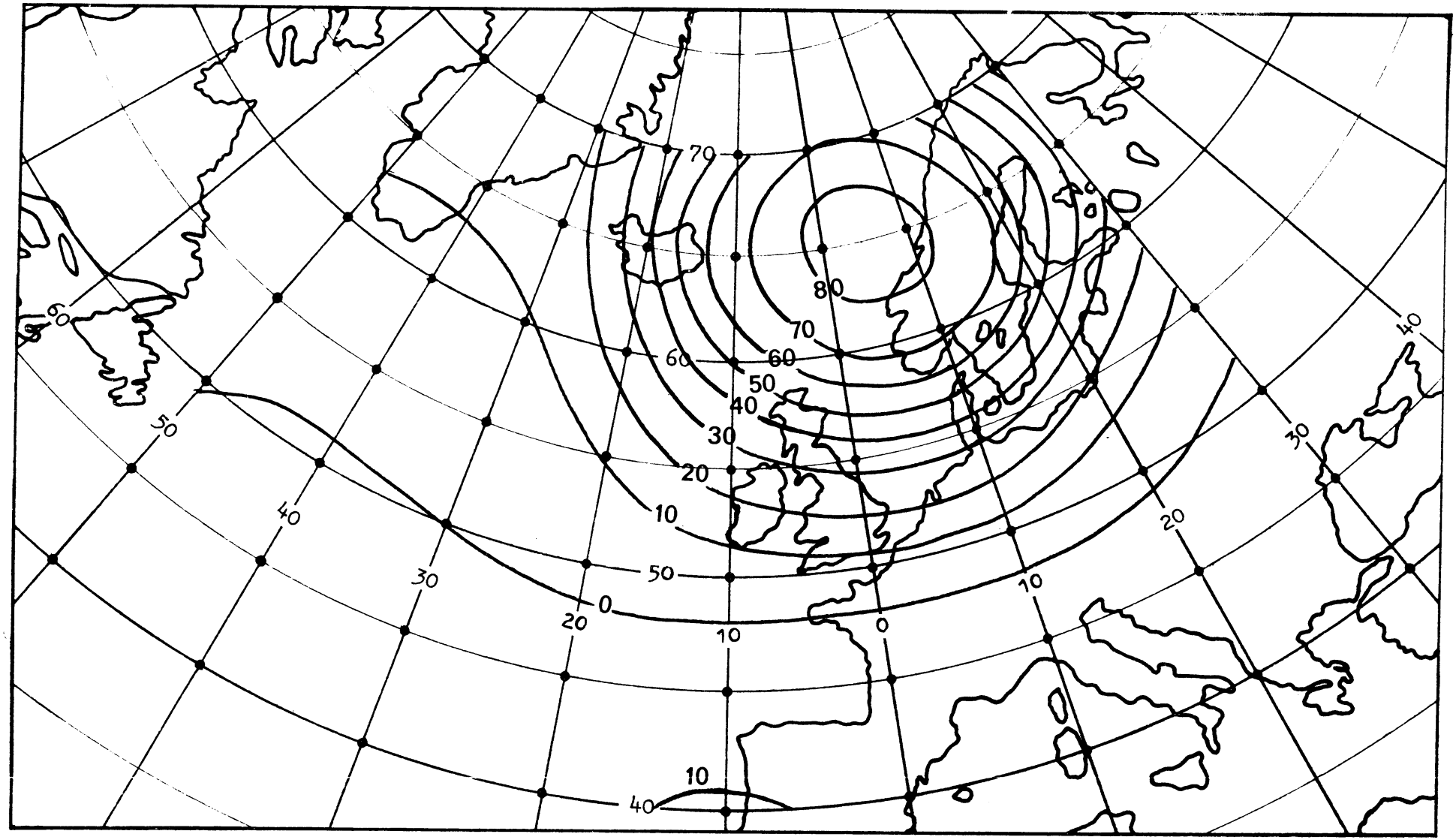

Fig. 3.18 Relatieve bijdrage tot variantie van de le eigenvector van augustus (in %).

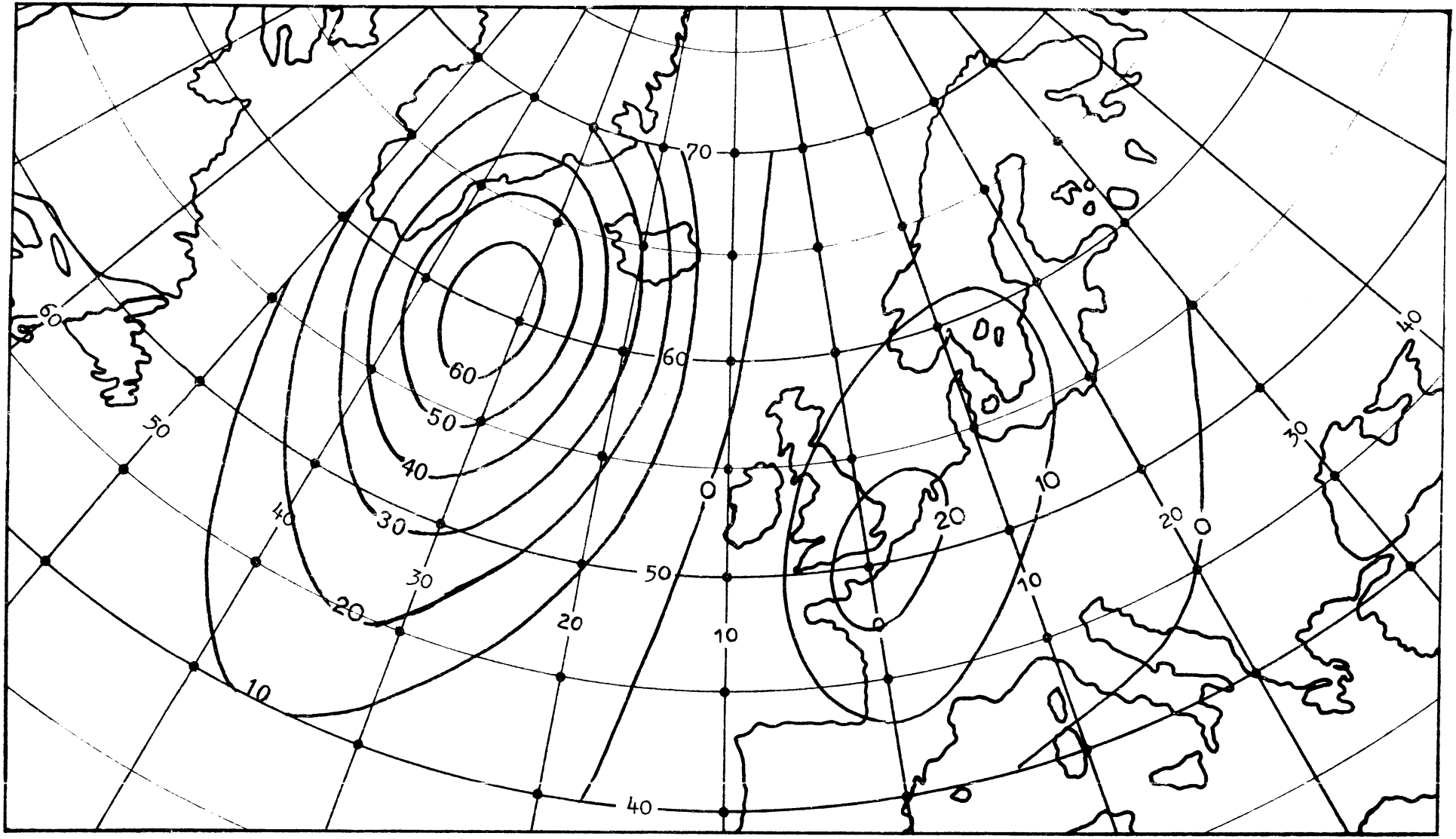

Fig. 3.19 Relatieve bijdrage tot variantie van de 2e eigenvector van augustus (in %).

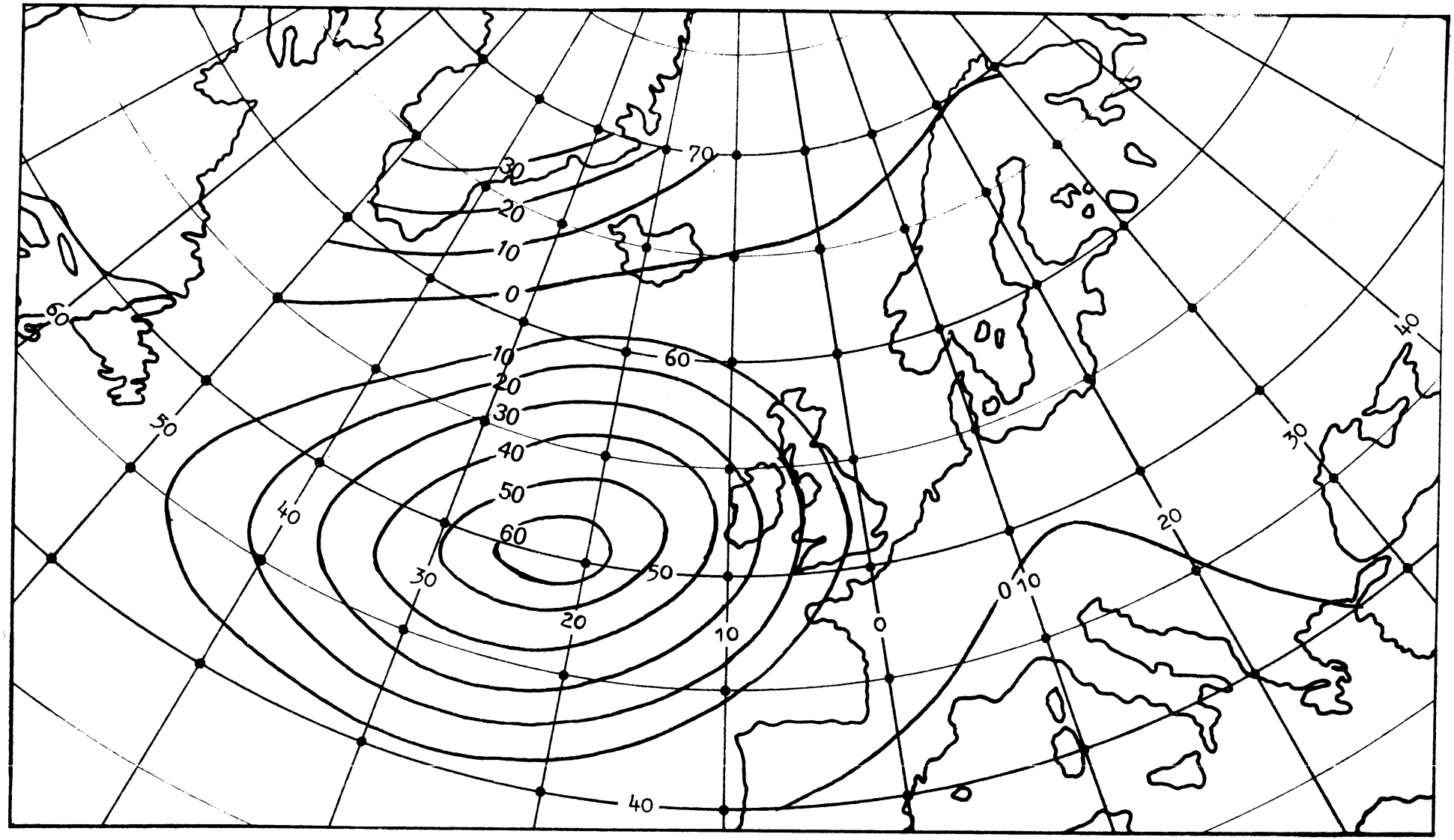

Fig. 3.20 Relatieve bijdrage tot variantie van de 3e eigenvector van augustus (in %).

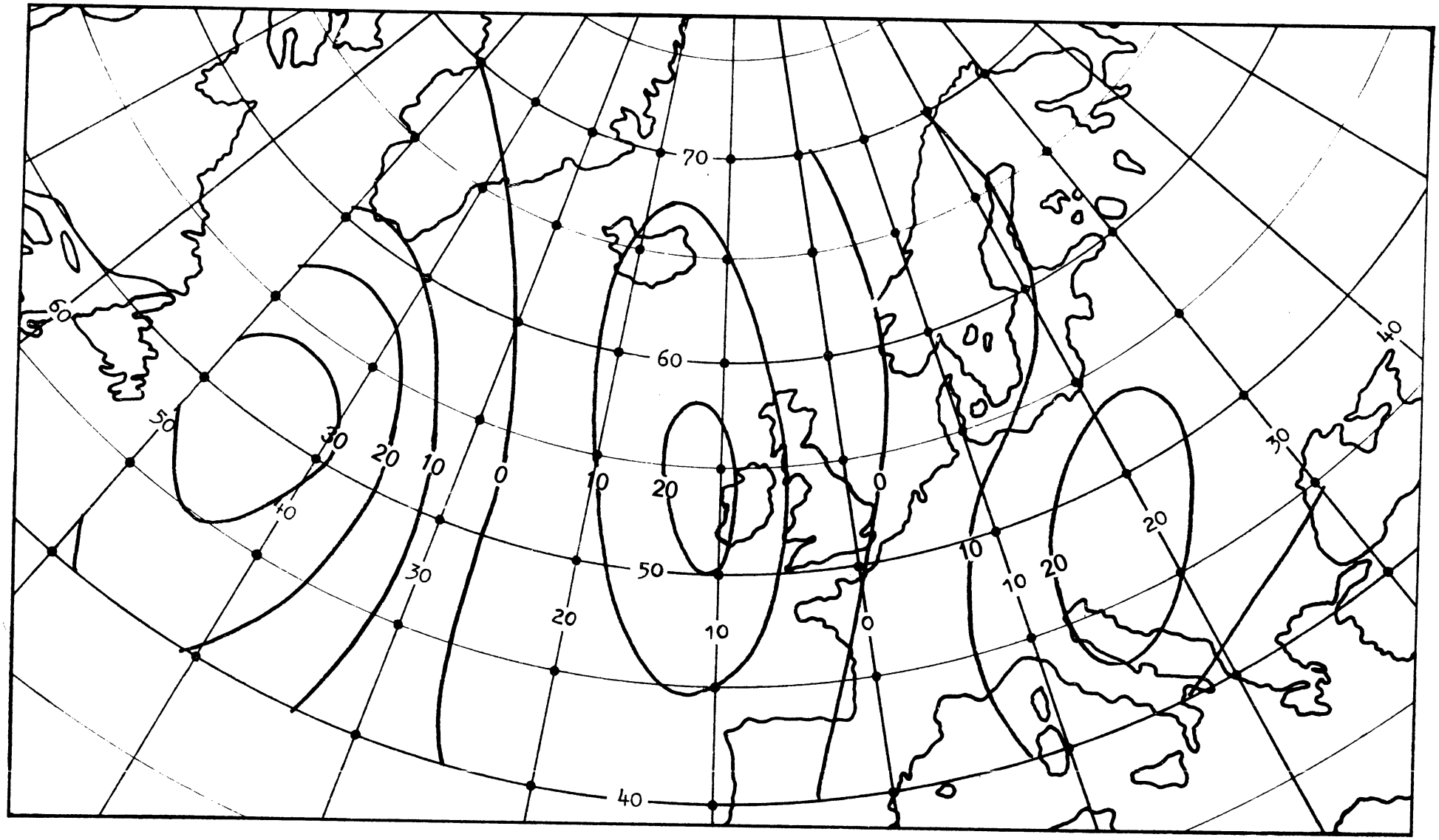

Fig. 3.21 Relatieve bijdrage tot variantie van de 4e eigenvector van augustus (in %).

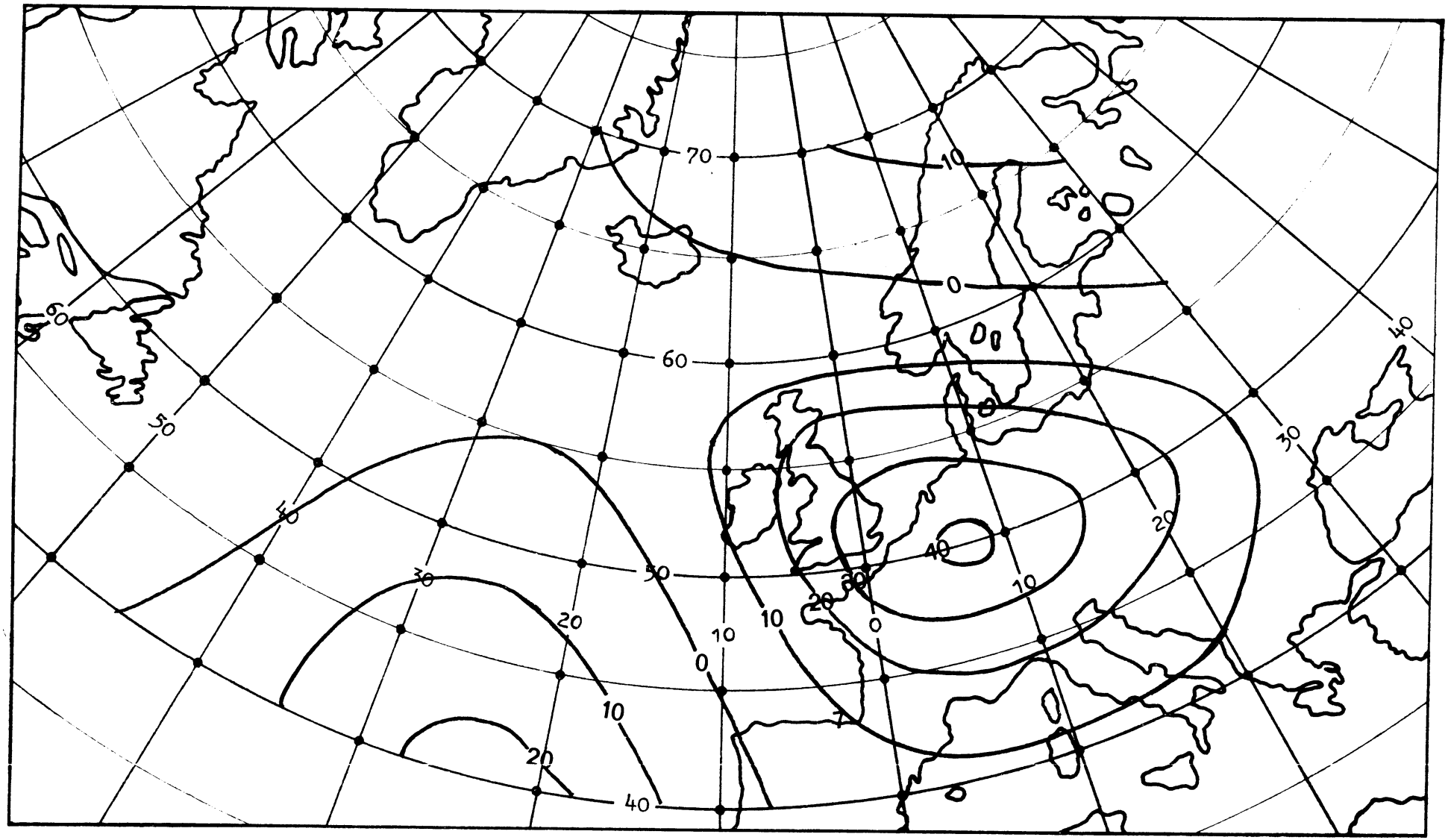

Fig. 3.22 Relatieve bijdrage tot variantie van de 5e eigenvector van augustus (in %).

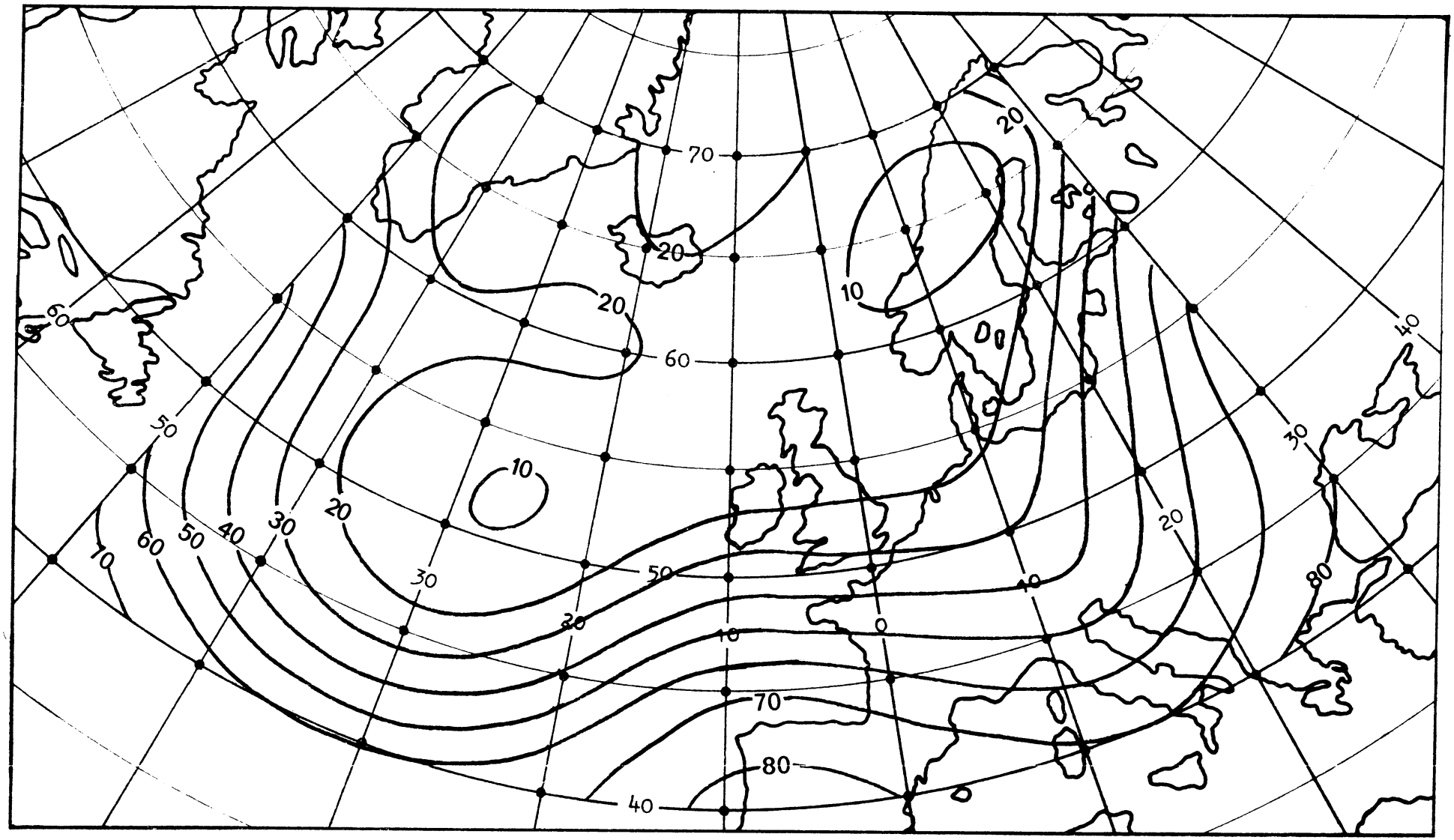

Fig. 3.23 Relatieve restvariantie na aftrek van 5 eigenvectoren (in %) voor augustus.

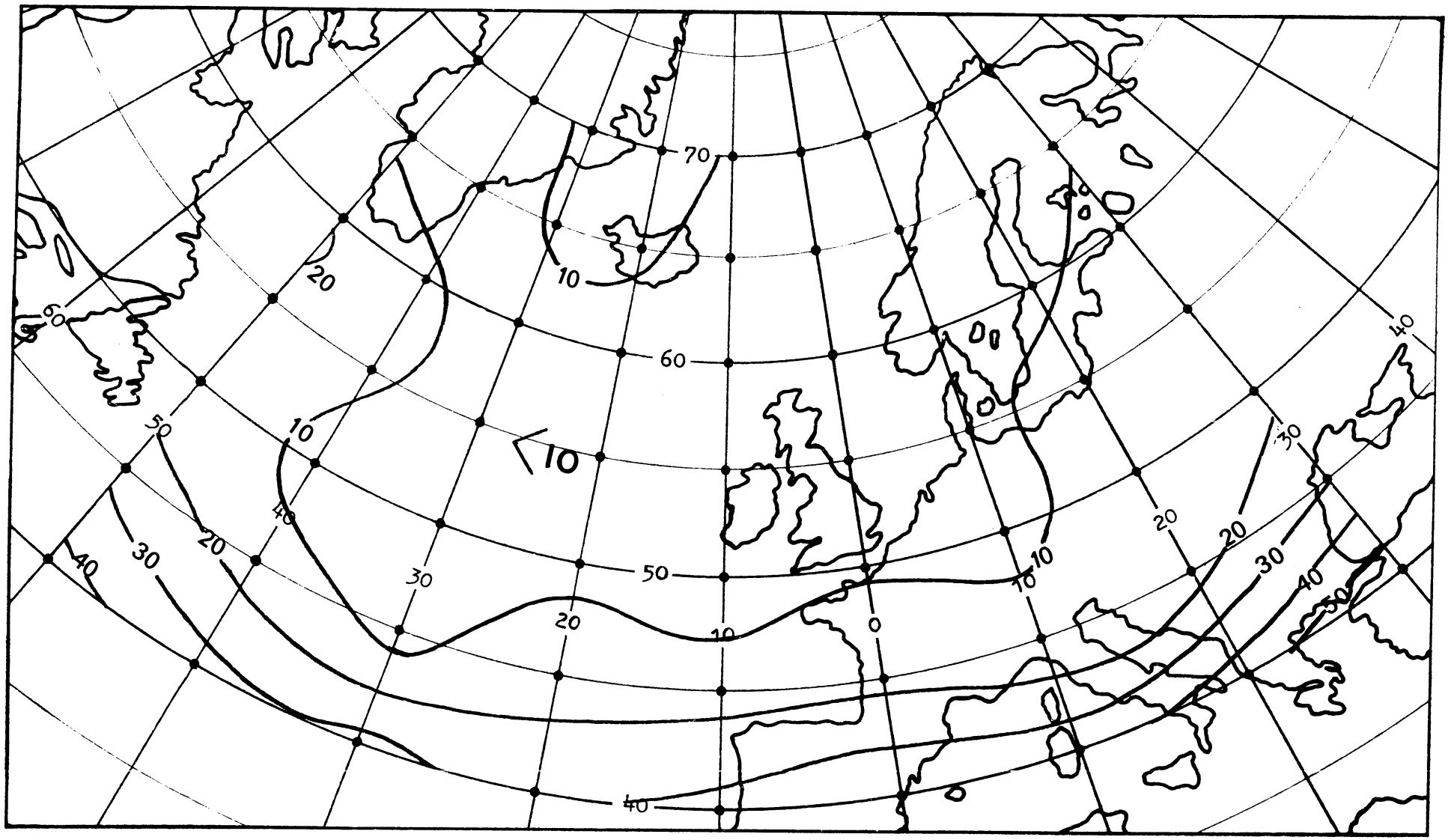

Fig. 3.24 Relatieve restvariantie na aftrek van 10 eigenvectoren (in %) voor augustus.

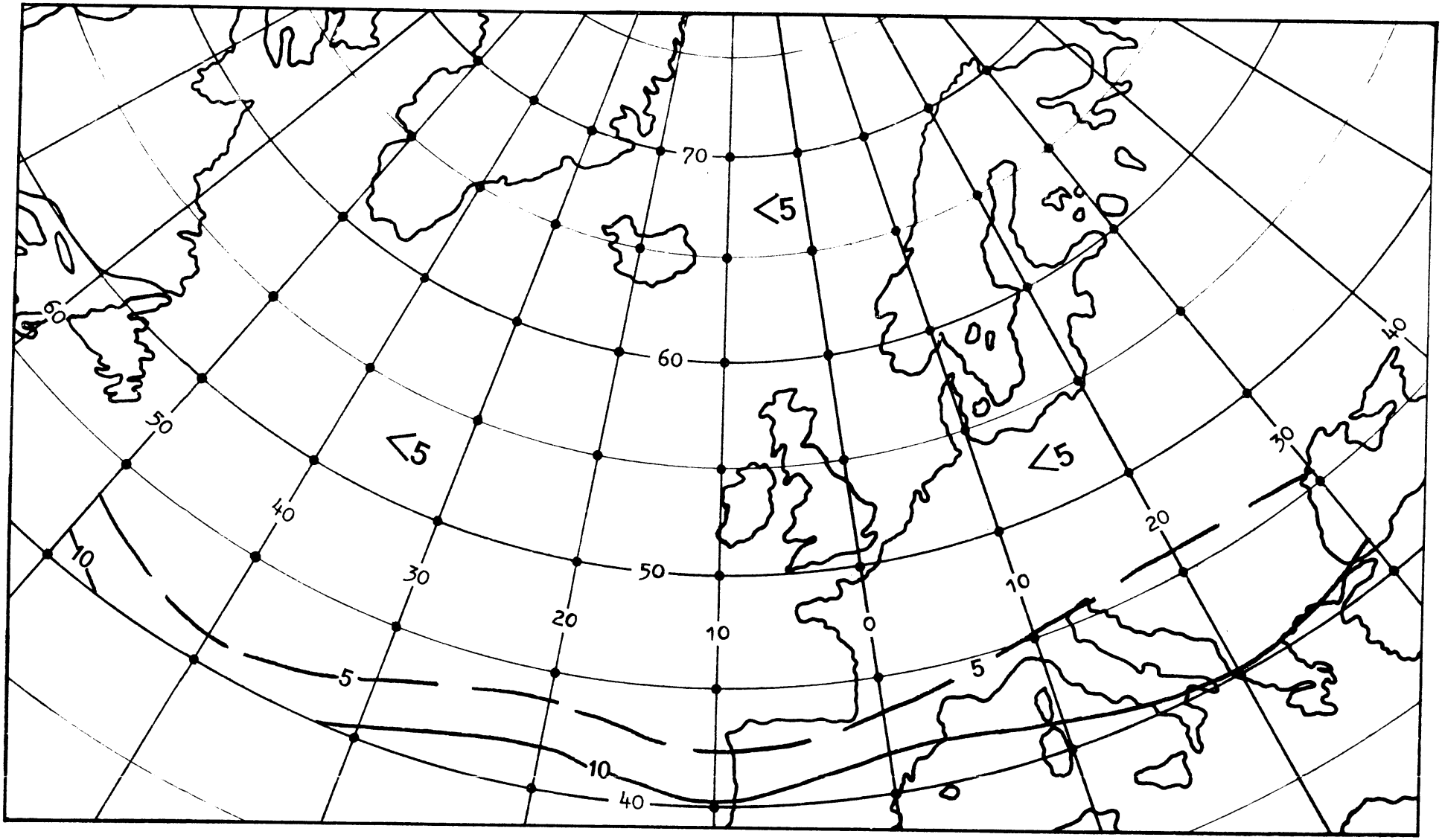

Fig. 3.25 Relatieve restvariantie na aftrek van 20 eigenvectoren (in %) voor augustus.

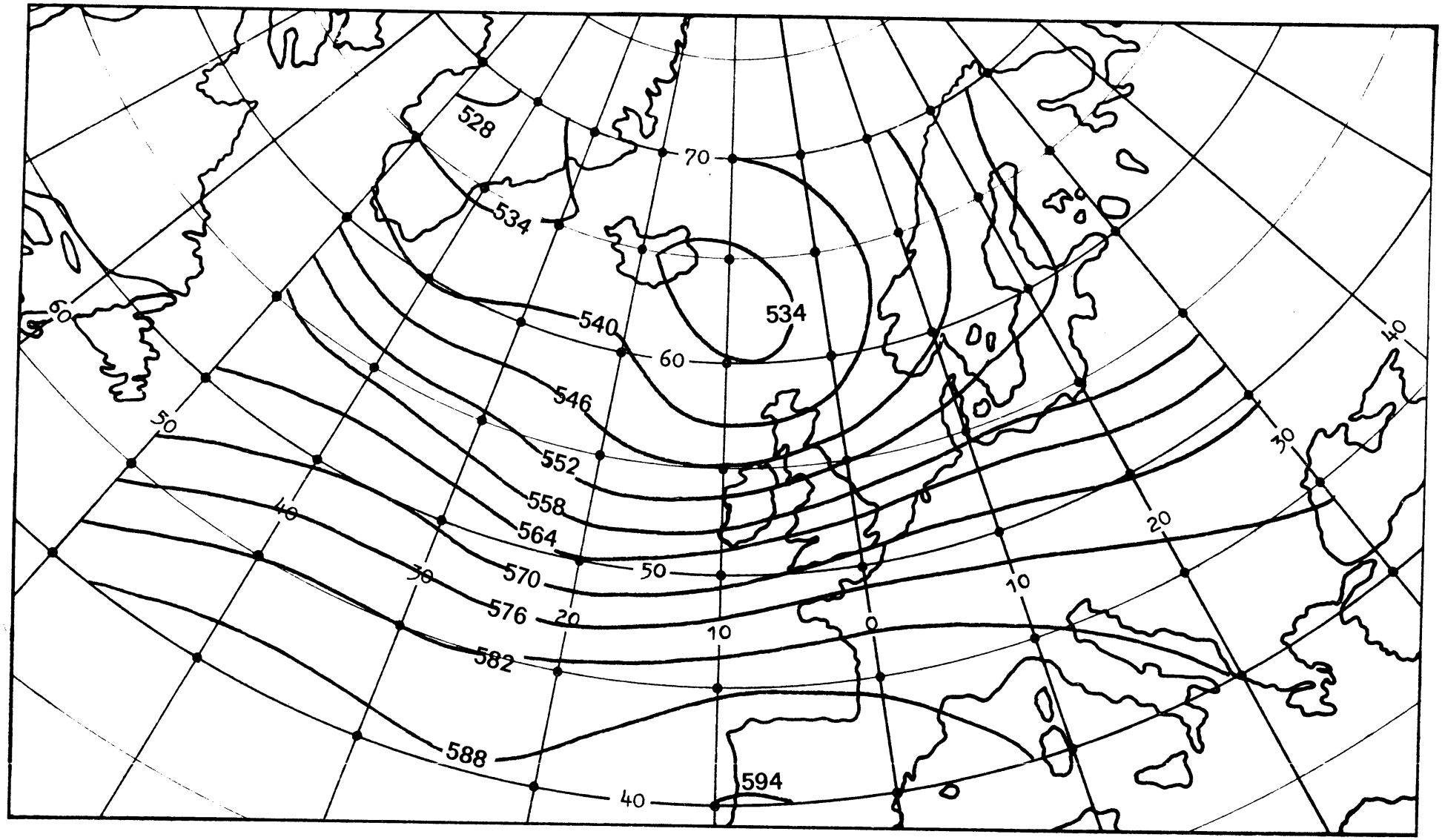

Fig. 3.26 Analyse 500 mbar-vlak 24-8-62.

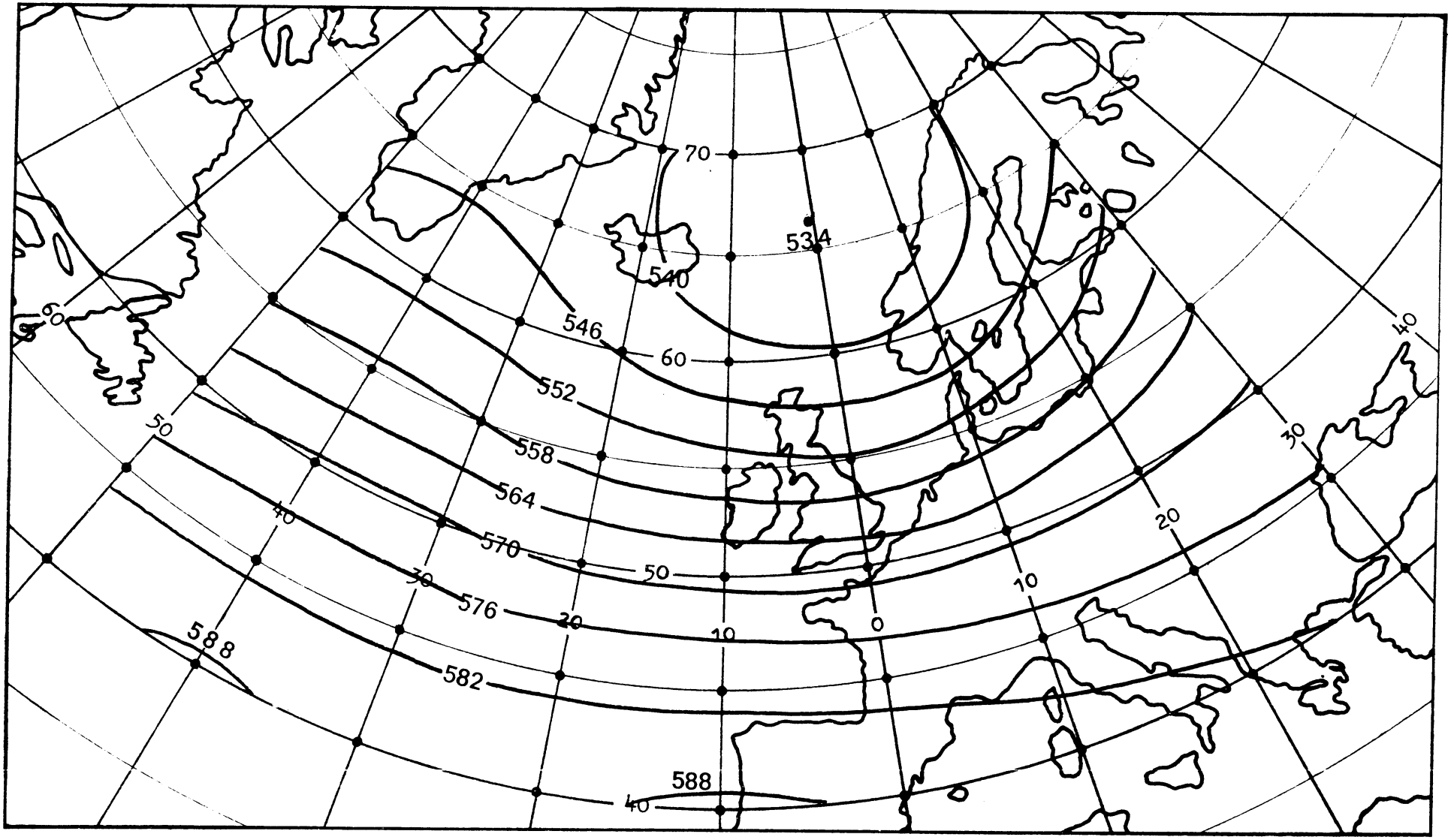

Fig. 3.27 Opbouw van het veld van 24-8-62 uit eigenvectoren:
Gem. + 1e eigenvector.

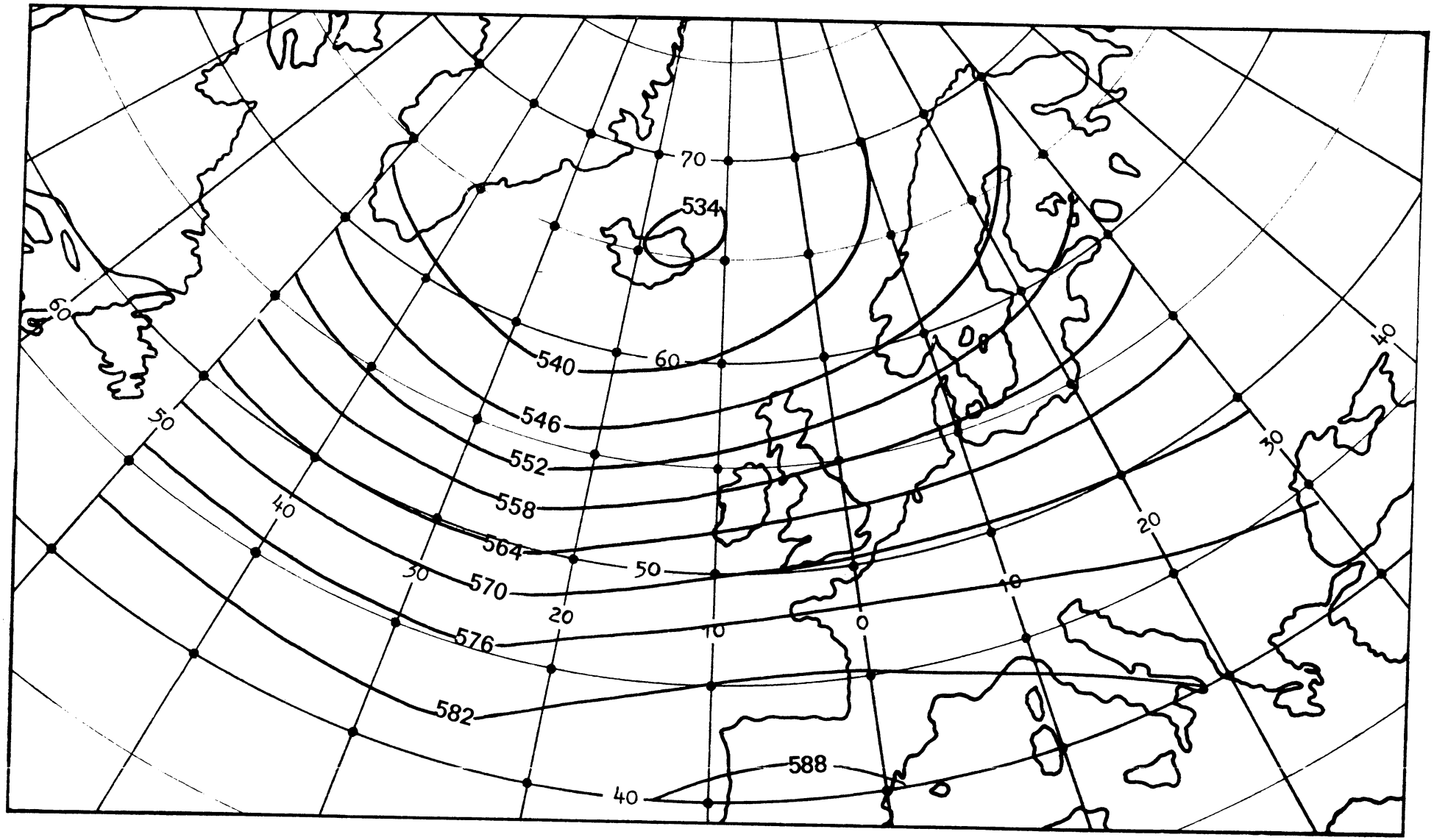

Fig. 3.28 Opbouw van het veld van 24-8-62 uit eigenvectoren:
Gem. + eigenvectoren 1 en 2.

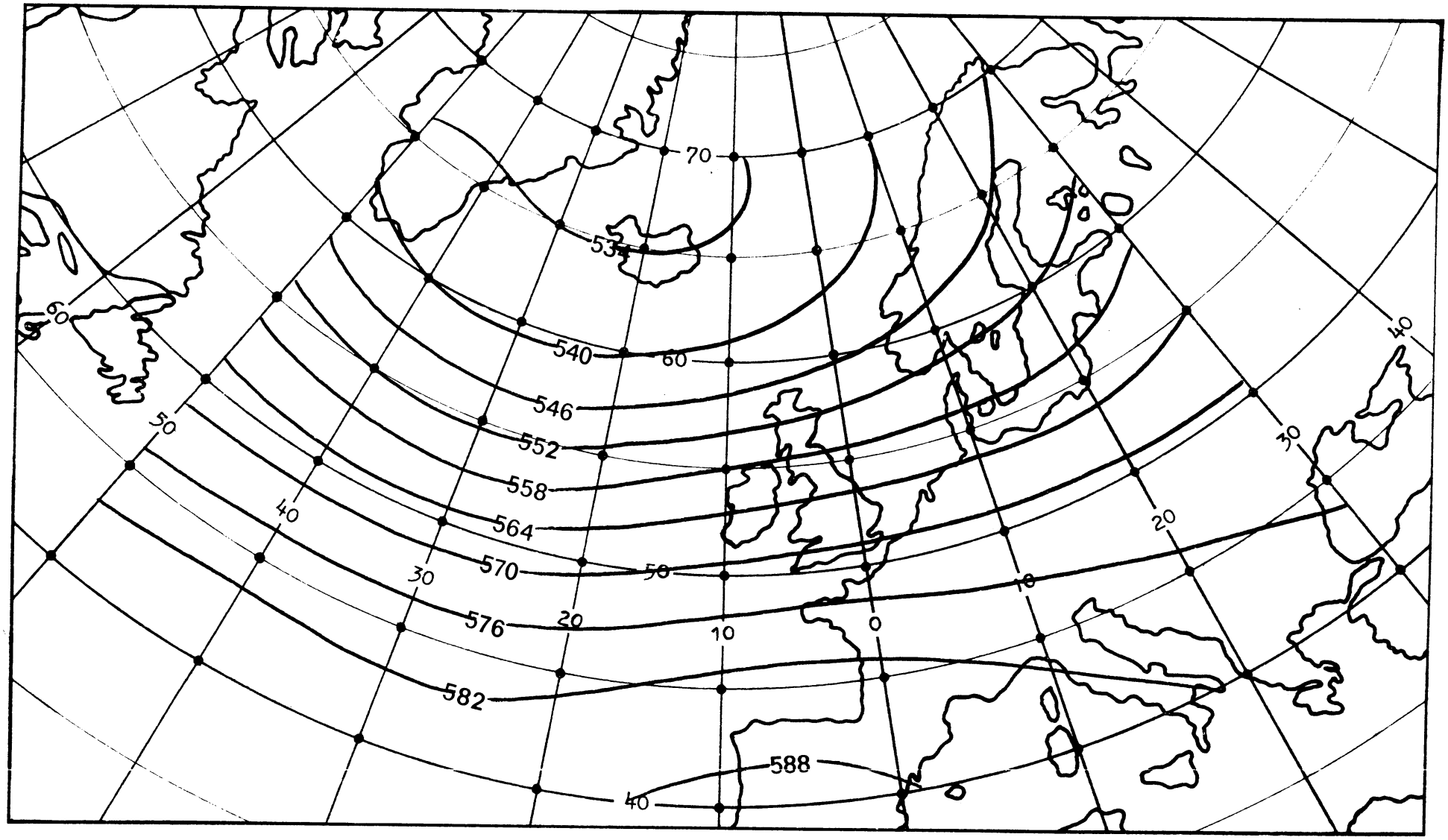

Fig. 3.29 Opbouw van het veld van 24-8-62 uit eigenvectoren:
Gem. + eigenvectoren 1 t/m 3.

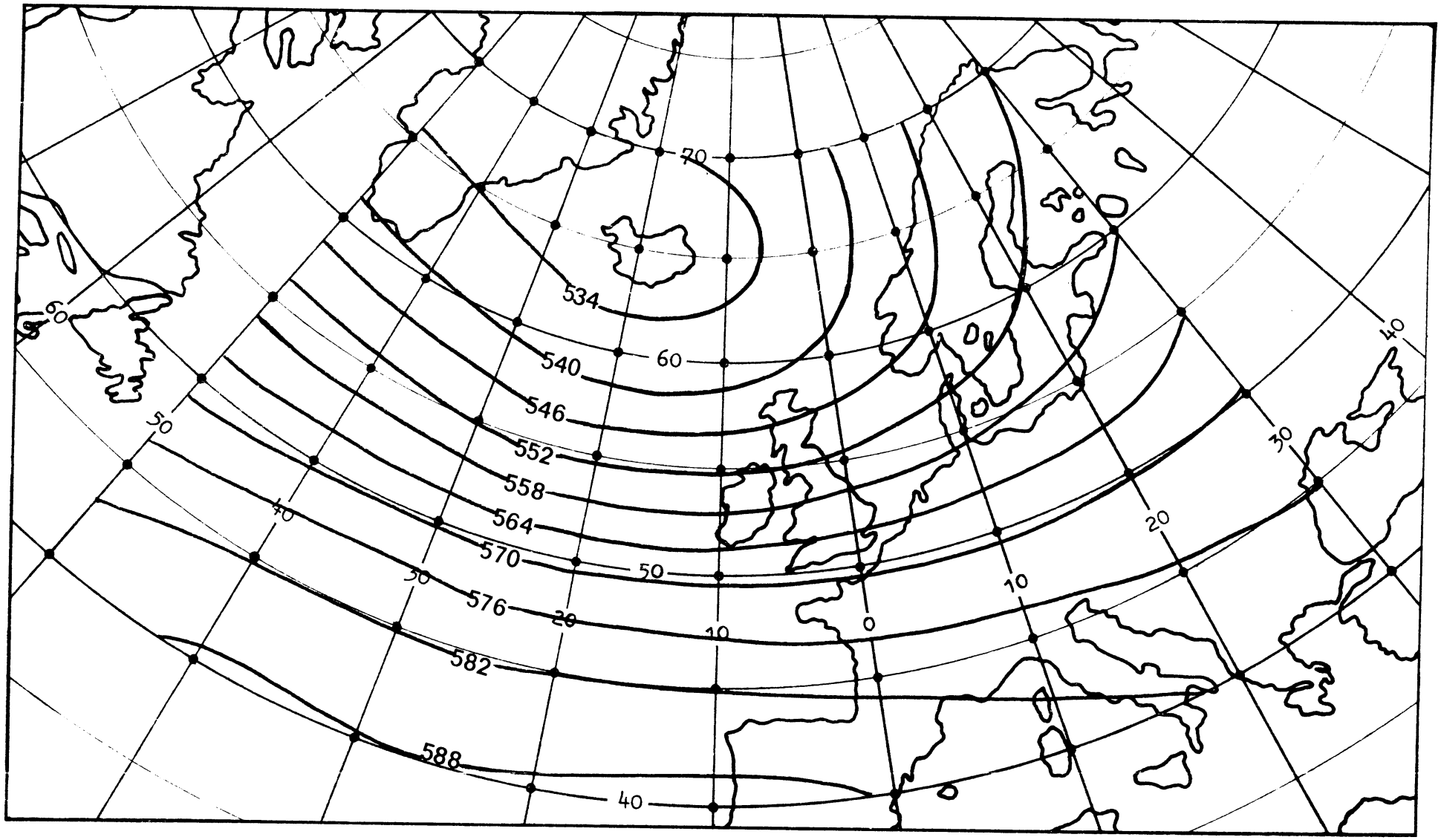

Fig. 3.30 Opbouw van het veld van 24-8-62 uit eigenvectoren:
Gem. + eigenvectoren 1 t/m 5.

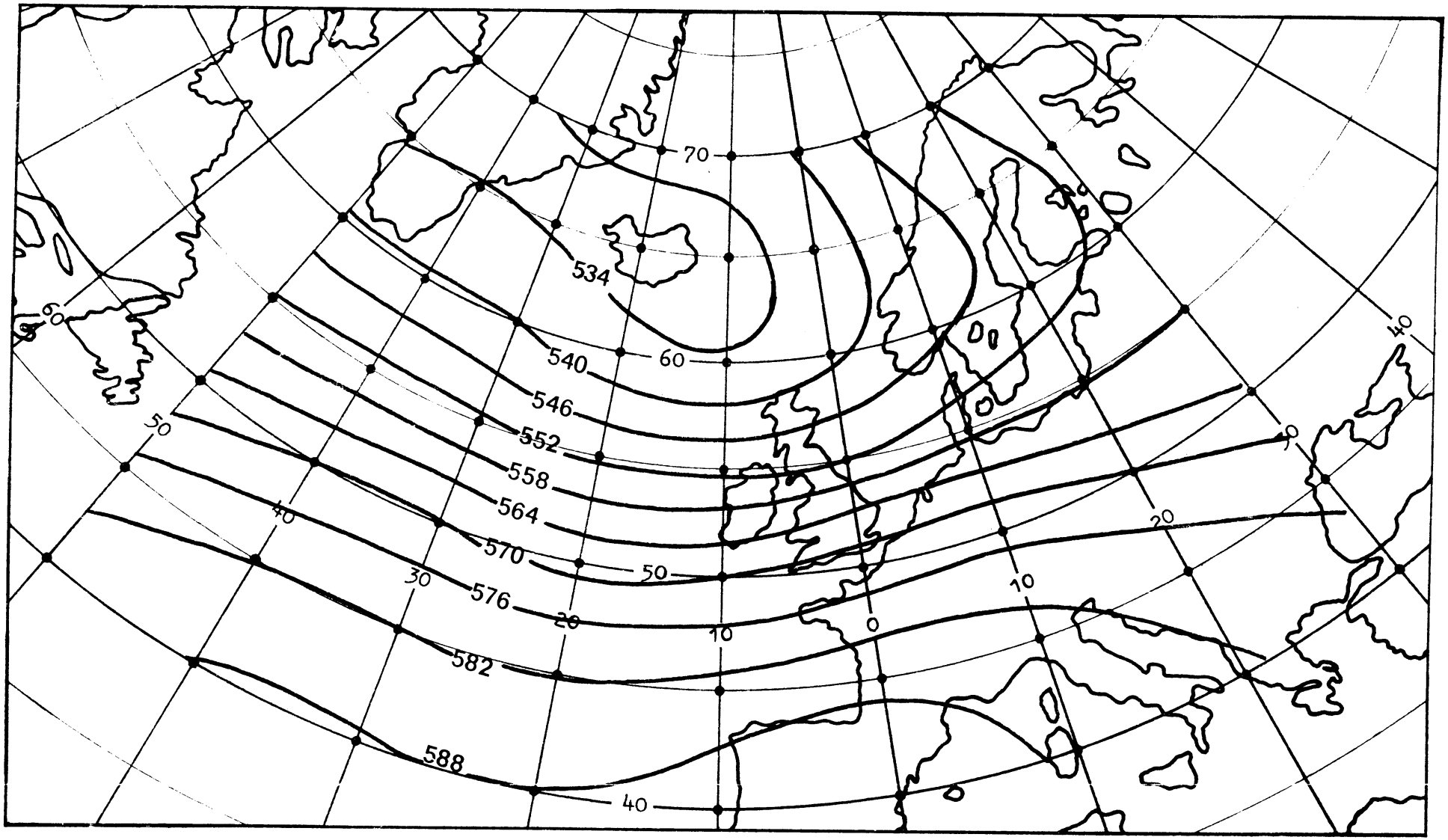

Fig. 3.31 Opbouw van het veld van 24-8-62 uit eigenvectoren:
Gem. + eigenvectoren 1 t/m 10.

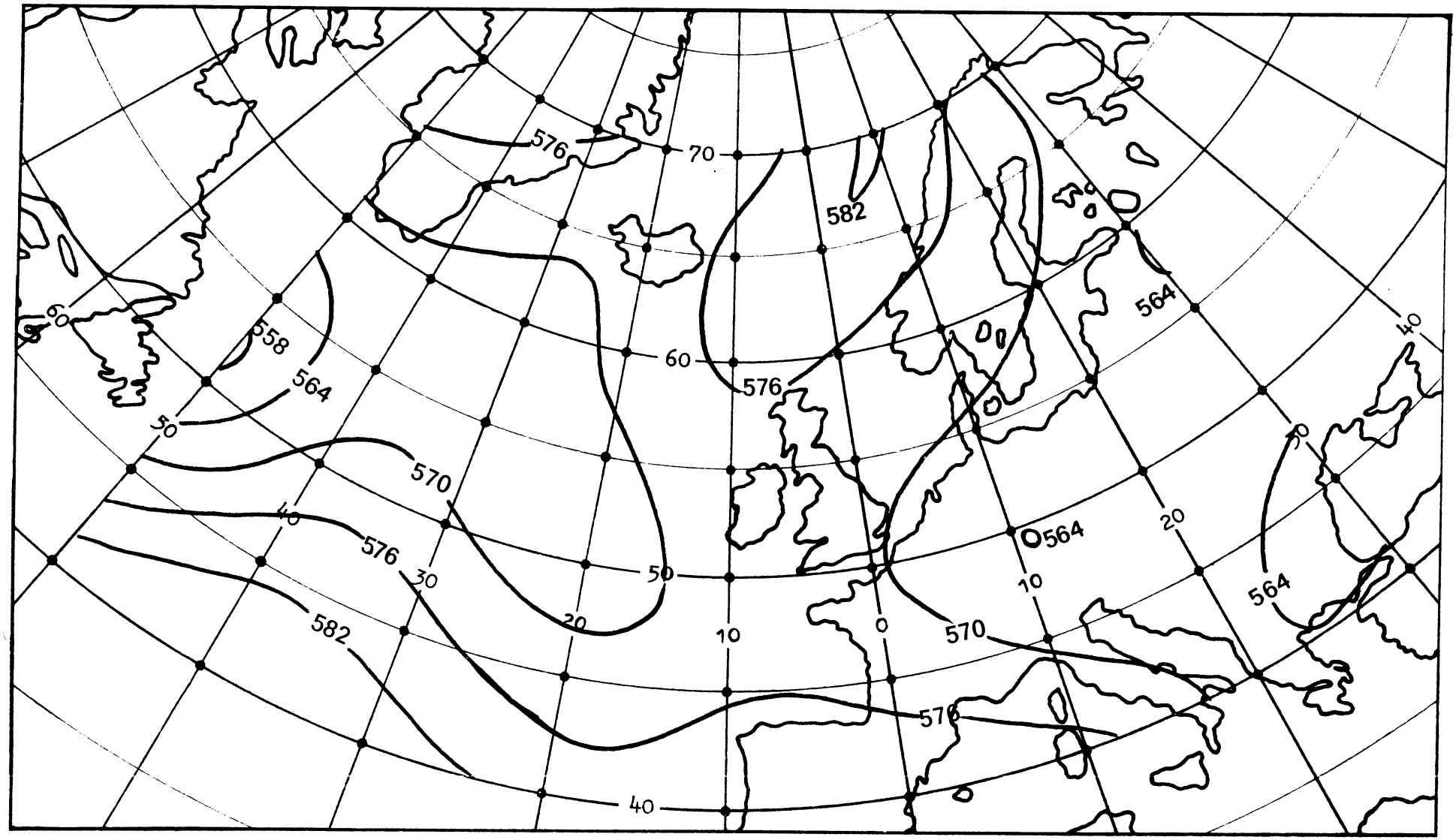

Fig. 3.32 Analyse 500 mbar-vlak 12-8-64.

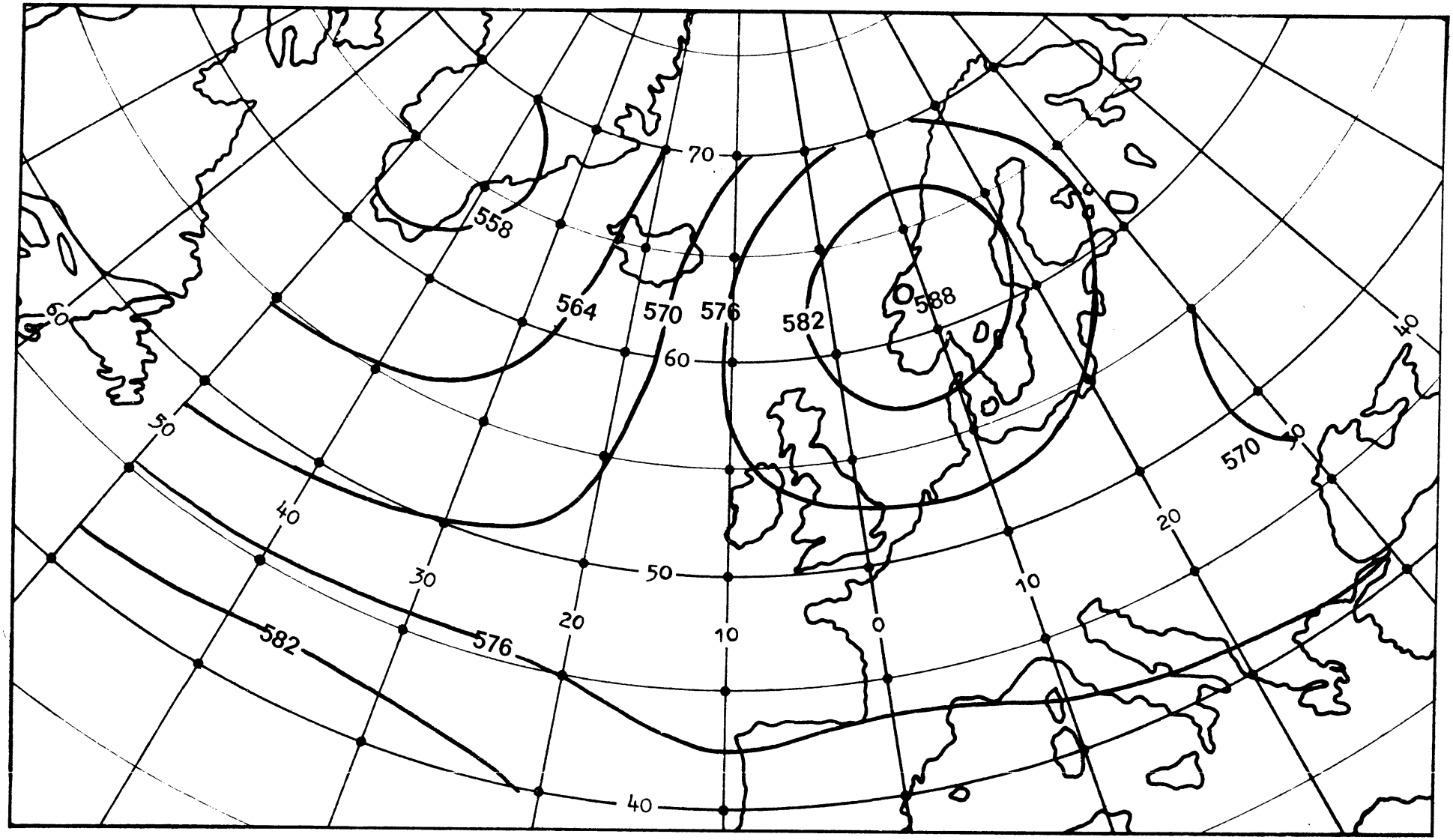

Fig. 3.33 Opbouw van het veld van 12-8-64 uit eigenvectoren:
Gem. + 1e eigenvector.

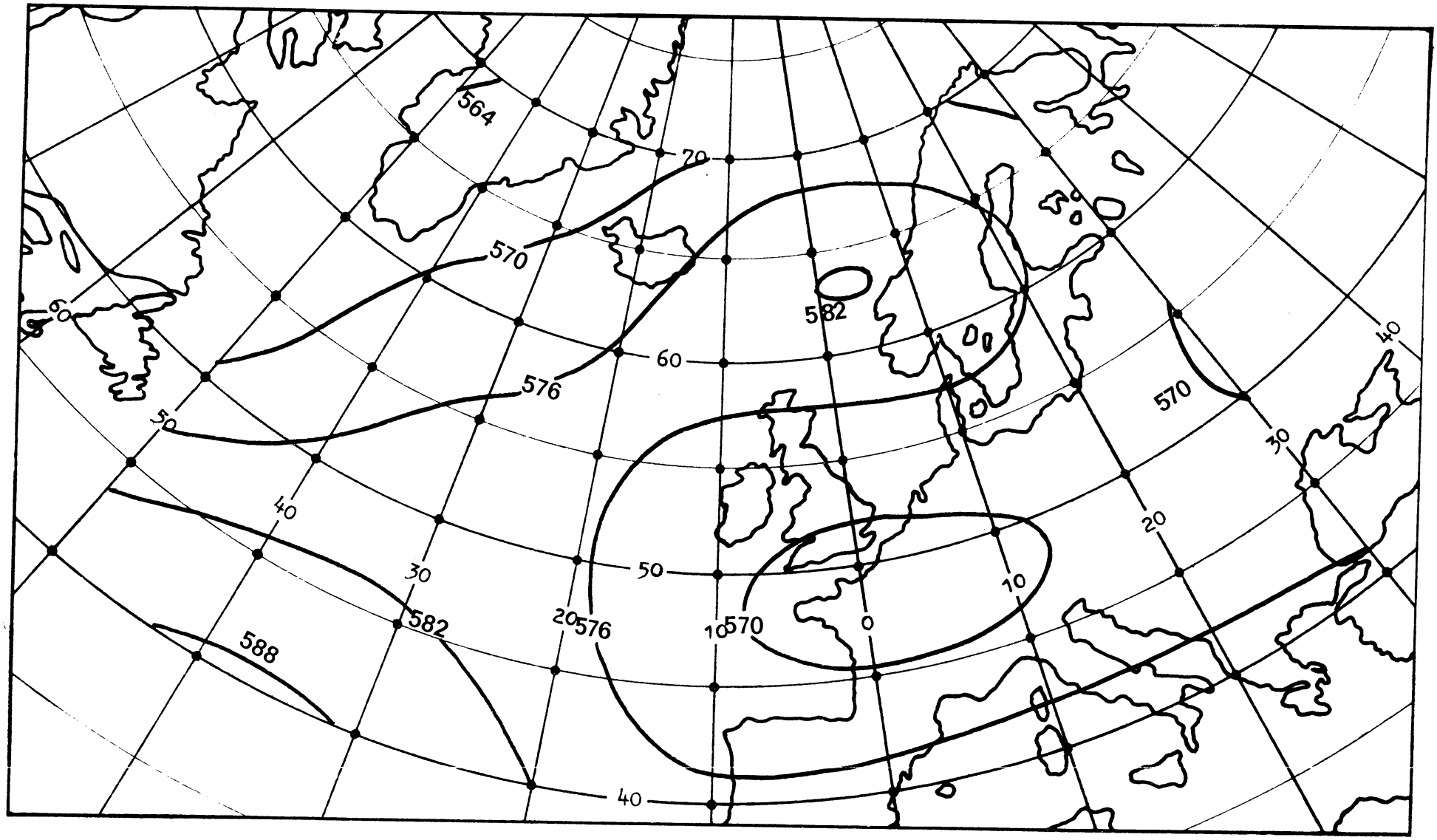

Fig. 3.34 Opbouw van het veld van 12-8-64 uit eigenvectoren:
Gem. + eigenvector 1 en 2.

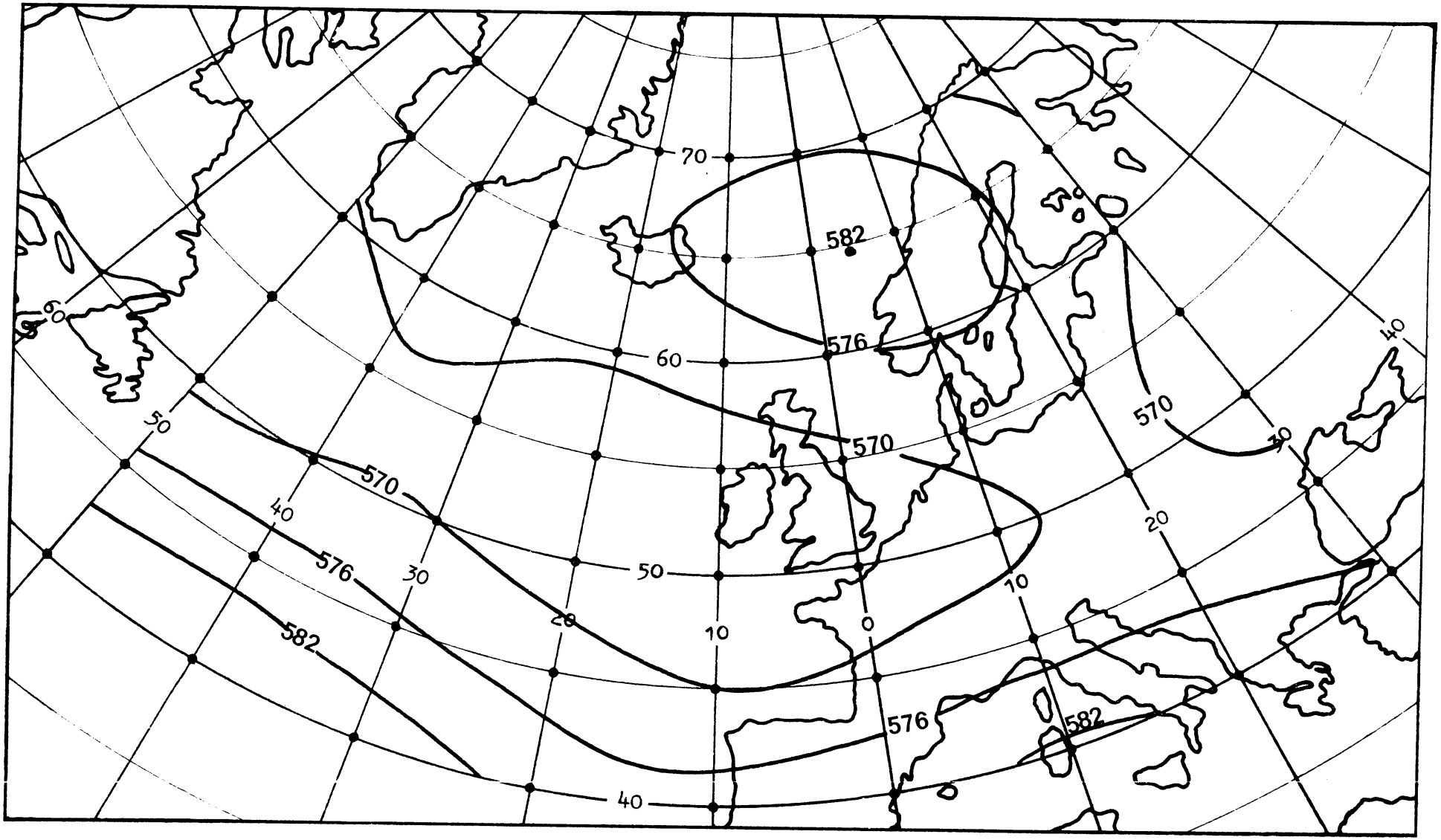

Fig. 3.35 Opbouw van het veld van 12-8-64 uit eigenvectoren:
Gem. + eigenvector 1 t/m 3.

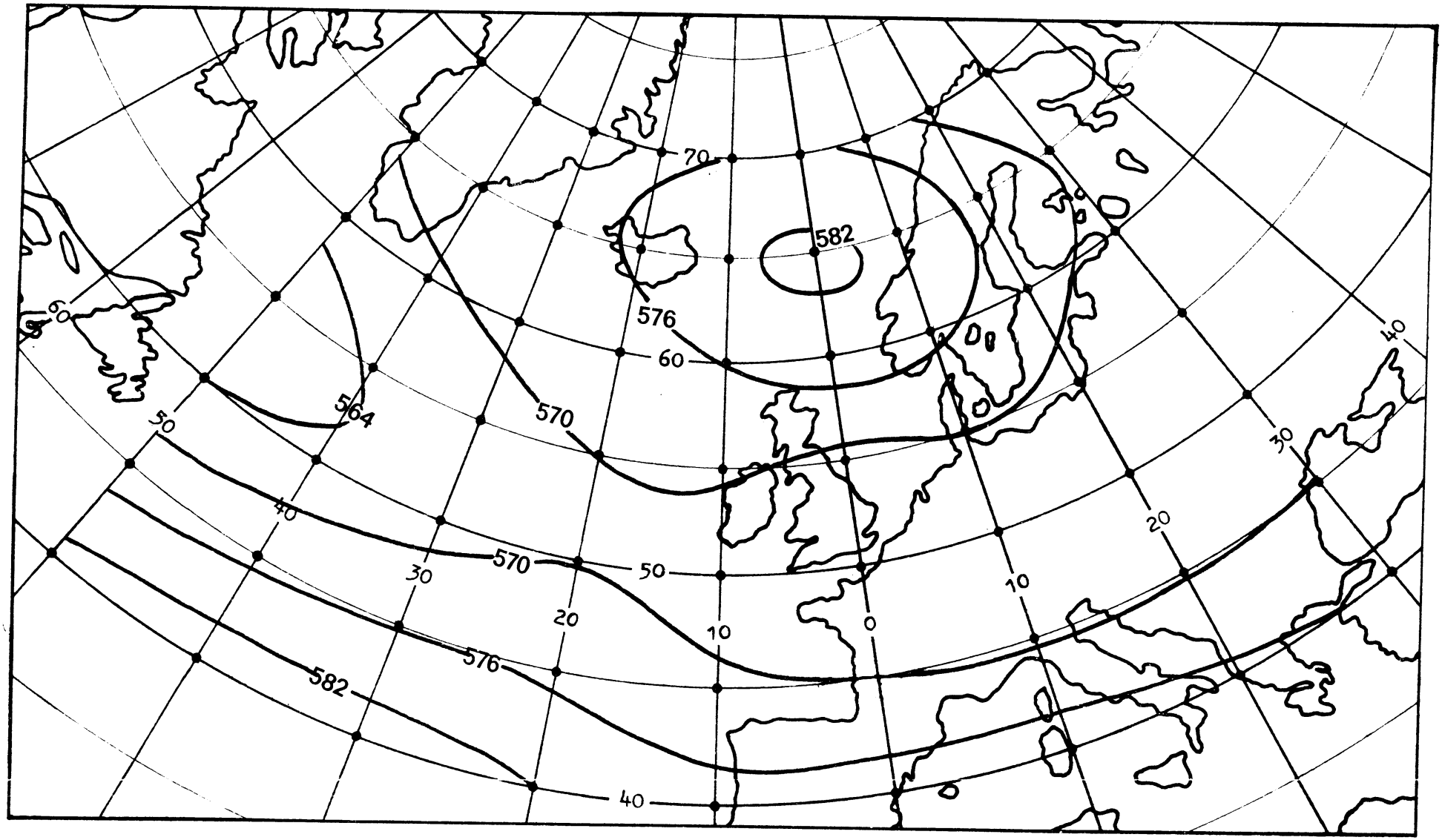

Fig. 3.36 Opbouw van het veld van 12-8-64 uit eigenvectoren:
Gem. + eigenvector 1 t/m 5.

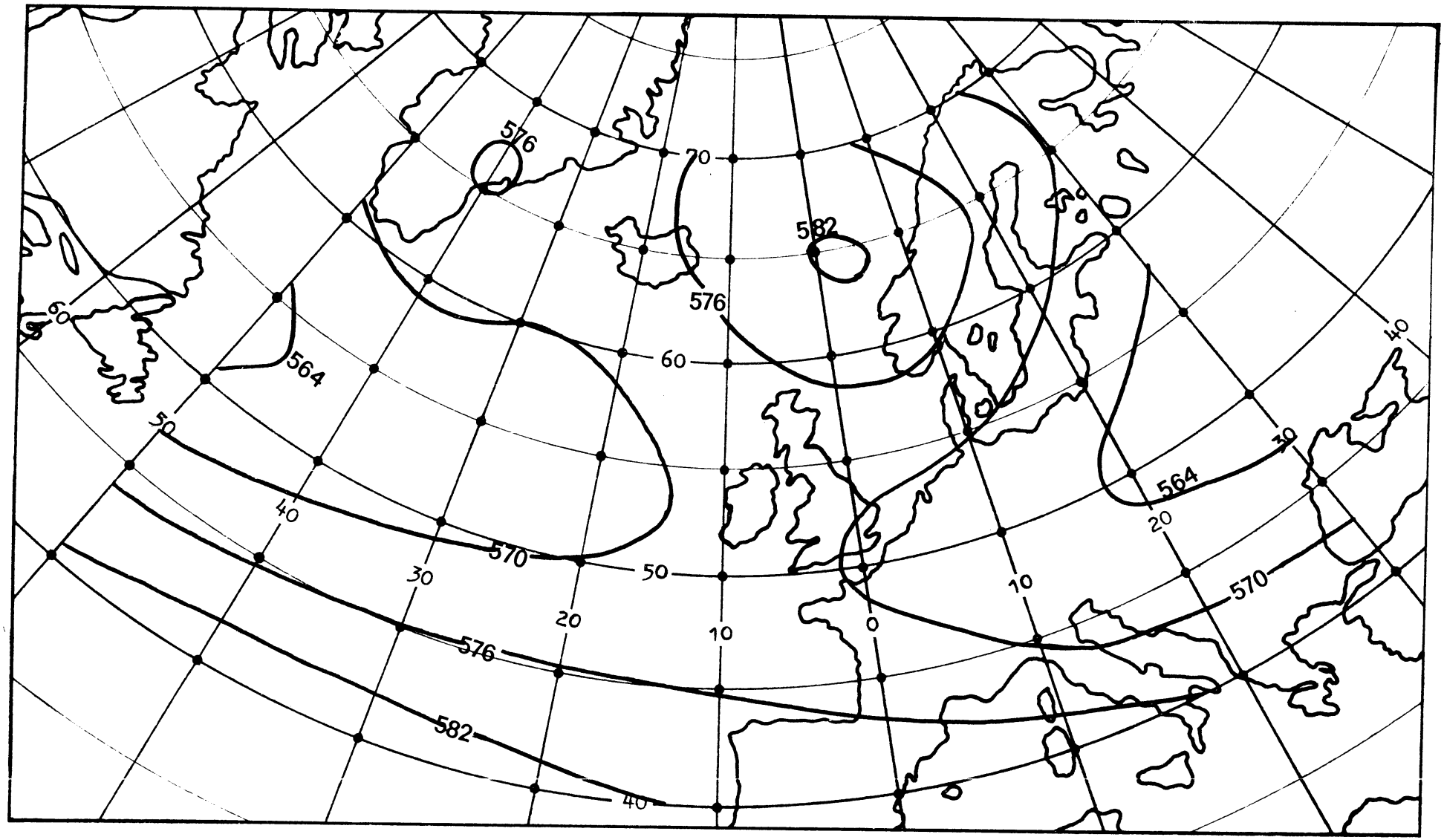

Fig. 3.37 Opbouw van het veld van 12-8-64 uit eigenvectoren:
Gem. + eigenvector 1 t/m 10.