A new method for verifying deterministic predictions of meteorological scalar fields

By J. VAN GALEN,[1] *Royal Netherlands Meteorological Institute, De Bilt*

(Manuscript received November 25, 1968; revised version April 2, 1969)

ABSTRACT

In numerical weather prediction most error statistics with respect to forecasts of air pressure systems suffer from an intrinsic blending of "shape" errors with location errors. Separation of both effects is desirable. A method is presented which appears to be suited to meet this objective as far as short-term predictions are concerned, neglecting possible rotational effects. The scheme has not yet been in use operationally, but the results of some test cases are discussed in this paper. No reference is made to probability predictions nor to any economic purposes of evalution from the users' point of view, since the only aim is the scientific one of checking the behaviour of numerical models.

1. Introduction

For many years attention has been given to the evaluation of weather forecasts. For a critical survey see Bleeker (1946). A general treatment of the problem may be found in the paper of Brier & Allen (1951). A review of the literature about this subject has been given by Johnson (1957). A more recent account of various measures of accuracy in forecasting with mathematical models of the atmosphere has been published by Holyoke (1965).

Referring in particular to probabilistic predictions, Murphy & Epstein (1967) distinguish between operational evaluation concerning "utility", i.e. the value of forecasts for the user, and empirical evaluation or verification in regard of the "perfection" or "accuracy" of predictions.

This paper deals with the verification of deterministic predictions of meteorological scalar fields with dynamical methods through numerical computation of values at gridpoints. The usual verification procedure of geopotential height predictions of some pressure-surface consists in calculating statistical quantities such as:

1. The correlation coefficient between the predicted height changes and those actually observed.
2. The root-mean-square height error.
3. The root-mean-square geostrophic vector wind error.

For a proper interpretation the results may be compared with scores yielded by some "blind" forecast such as a climatological or a "persistence" forecast. In this way a verification program can provide for each forecast or a number of forecasts some "figures of merit" (Holyoke, 1965).

Such figures tell us that one model "works better" than the other. But in which respect does it work better? One cannot get a clear notion about the answer unless by visual, and therefore subjective, inspection. Obviously, all detailed information has been compressed into just one handy figure. For example, not a single indication is given about the spatial distribution of errors.

In order to avoid this difficulty one sometimes prepares scalar error fields. Daily inspection of error maps may be advisable, otherwise one feels inclined to reduce such (time-averaged) error charts to a single objective statistical measure.

A disadvantage of most scalar statistics is that the scalar error measure may be seriously affected by a superimposed long wave error (or a constant), whereas the—from the forecaster's point of view much more important—predicted gradients behave fairly well. This is a spectral problem that could be solved by decomposing

[1] Present affiliation: IBM, Development Laboratory, Uithoorn, Netherlands.

the geopotential height fields into their Fourier-components. But wave number statistics are produced by spatial averaging, so that no error information can be provided for certain areas as distinct from other areas. Another aspect is the combined effect on both the scalar and vector error measures of the displacement and of the development of air pressure systems.

The purpose of this paper is to introduce a separation method for both these effects, so that location and shape can be verified independently. The method might be said to produce two coupled "fields of figures of merit": one scalar field representing the areal distribution of the "goodness of shape" (σ), and one location error vector (ϵ) field. A third field might be added to represent scalar deviations in geopotential height. This is not an error chart in the ordinary sense, but one from which displacement errors have been eliminated.

2. Outline of the method

Rarely the observed and predicted positions of depressions, troughs etc., coincide and therefore the underlying thought of the verification method to be developed here is to fix a system in the verifying chart first and then to search in the forecast chart for the system that shows the greatest geometrical resemblance to the reference system. In doing so some "pattern recognition" technique has to be applied.

For the purpose of this study it is convenient to define a system, or a pattern, as the information contained in an arbitrary rectangle of $m \times n$ gridpoints. An interpolated system is defined as an equally-sized rectangular part of the continuous field, though approximated by interpolation from gridpoint values. The size of the rectangle is assumed to be constant during each verification, but remains for the moment unspecified.

Next, some measure for "pattern similarity" must be provided, preferably ranging from -1 to $+1$. For trial attempt it was decided to work with the linear correlation coefficient (ϱ) as a similarity coefficient (s.c.).

Now the observed system, once fixed, can be correlated with the predicted system having the same location. The s.c. may be thought of as stored in the geometrical centre of the predicted system. Furthermore, the observed system is correlated with predicted systems in the neighbourhood, with the s.c. stored in their respective centres. This is done until a maximum value (ϱ_m) has been found, using a sub-gridscale interpolation procedure (see Appendix A). In this way a vector is defined pointing from the position corresponding with the centre of the observed system to the position of the centre of the "best" interpolated system in the forecast chart. This vector may be interpreted as the local "location error" ϵ. The value ϱ_m represents by definition the "goodness of shape", irrespective of location, in short: $\sigma \overset{\mathrm{DEF}}{=\!=} \varrho_m$.

Theoretically σ may have the value $+1$, which means that of the observed system, though maybe not at the right place, the direction of the winds was perfectly predicted. On the other hand any zero vector ϵ indicates a correct location, no matter how high, or low, the degree of resemblance σ may be.

The process is to be repeated for all other observed systems within the verification area. The ϵ-field obtained in this way and isolines drawn for σ, provide the areal distributions of location errors and shape errors, respectively.

The method may be illustrated by the following example. Let some mathematically idealized low in the topography of the 500 mbar pressure-surface be given by the circles

$$(x-\zeta)^2 + (y-\eta-\mu_k)^2 = (\mu_k+\nu_k)^2 \qquad (1)$$

where

x, y are Cartesian grid coordinates,
ζ, η are constants,
μ_k, ν_k are parameters with index k.

The parameter μ_k defines with η the position of the centre point of each circle on the line $x=\zeta$ and also its radius together with ν_k. μ_k is suitably chosen as

$$\mu_k = k \cdot \tan\left(\frac{k\pi}{2l}\right), \quad 0 \leqslant k < l \qquad (2)$$

where l is taken as a constant.

If k is related to the geopotential height, for instance by

$$z = z_0 + 4k \text{ (geopotential decameters)} \qquad (3)$$

where z_0 is a constant, then with any chosen ν_k from (1), (2) and (3) several different situations may be defined. Suppose that an "observed"

Table 1

	ζ	η	ν_k	l	z_0
OBS	0	0	$\sqrt{2k}$	15	500
PRD	$+1.2$	-0.7	$\sqrt[3]{\frac{1}{2}k}$	18	480

situation (OBS) is given by $\zeta = \eta = 0$, $\nu_k = \sqrt{2k}$, $l = 15$, $z_0 = 500$, and that a "predicted" situation (PRD) shows the values $\zeta = +1.2$, $\eta = -0.7$, $\nu_k = \sqrt[3]{\frac{1}{2}k}$, $l = 18$, $z_0 = 480$ (see Table 1).
For $|x|, |y| \leqslant 5$ these two situations are displayed in Fig. 1, where the isohypses correspond to allowable integer values of k (see (2)).

It will be clear that any reasonable checking must take into consideration:

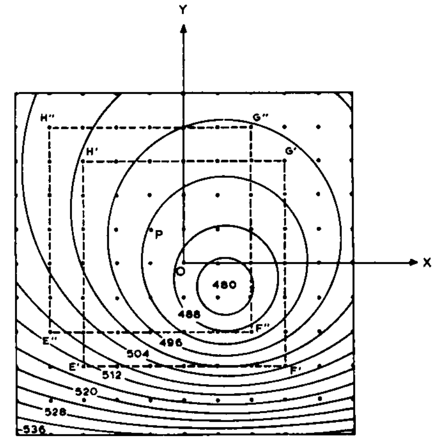
I. The *location shift* (due to ζ and η).

II. The *shape disparity* (associated with ν_k and l).

III. The *difference in height* at the centre (caused by z_0).

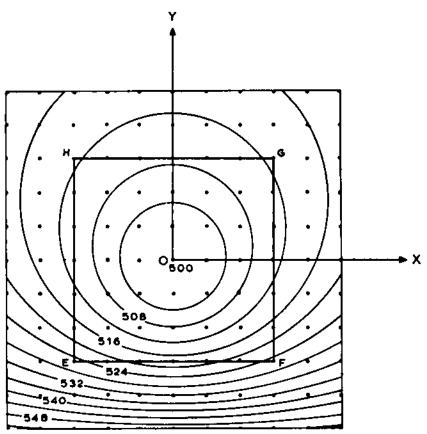
Possible rotational effects are neglected throughout this paper.

Taking for the dimensions of the rectangle $m = n = 7$, let the central part of the OBS-low, given by the values at the 49 gridpoints of square EFGH in Fig. 1*a*, constitute the reference system. Then with respect to this system the s.c. for each possible PRD-system within the area $|x|, |y| \leqslant 5$ (always of course with $m = n = 7$) is calculated. So, for example, the systems $E'F'G'H'$ and $E''F''G''H''$ (Fig. 1*b*) show a s.c. of $+0.766$ and $+0.224$ respectively, stored in their geometrical centres O and P (Figs. 1*b* and 2). From Fig. 2 it is seen that Q is the gridpoint with the greatest s.c. and therefore the interpolation procedure (Appendix A) is applied to the surrounding square KLMN. From this, σ is found to be $+0.984$ in point A' with grid coordinates ($+1.27$, -0.48), whereas the true position of the centre of the PRD-low



b

Fig. 1 b. Mathematical example 500 mbar contours. "Predicted" situation (PRD) with systems $E'F'G'H'$ and $E''F''G''H''$.



a

Fig. 1 a. Mathematical example 500 mbar contours. "Observed" situation (OBS) with reference system *EFGH*.

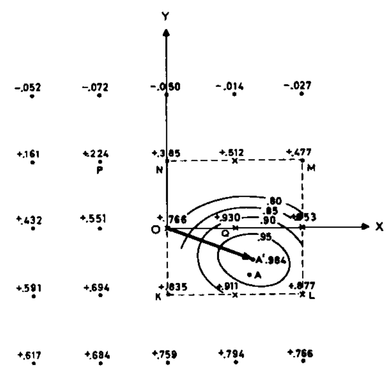


Fig. 2. Values for the similarity coefficient (s.c.) with interpolated maximum value ($\varrho_m = +0.984$) at point A'. The isolines represent equal values of the s.c. according to the polynomial fitting of the values labelled with an x-symbol, as discussed in Appendix A. The vector $\mathbf{OA}'$ is thought of as the location error, whereas the "goodness of shape" is measured by ϱ_m. All this refers to the system *EFGH* shown in Fig. 1 *a*.

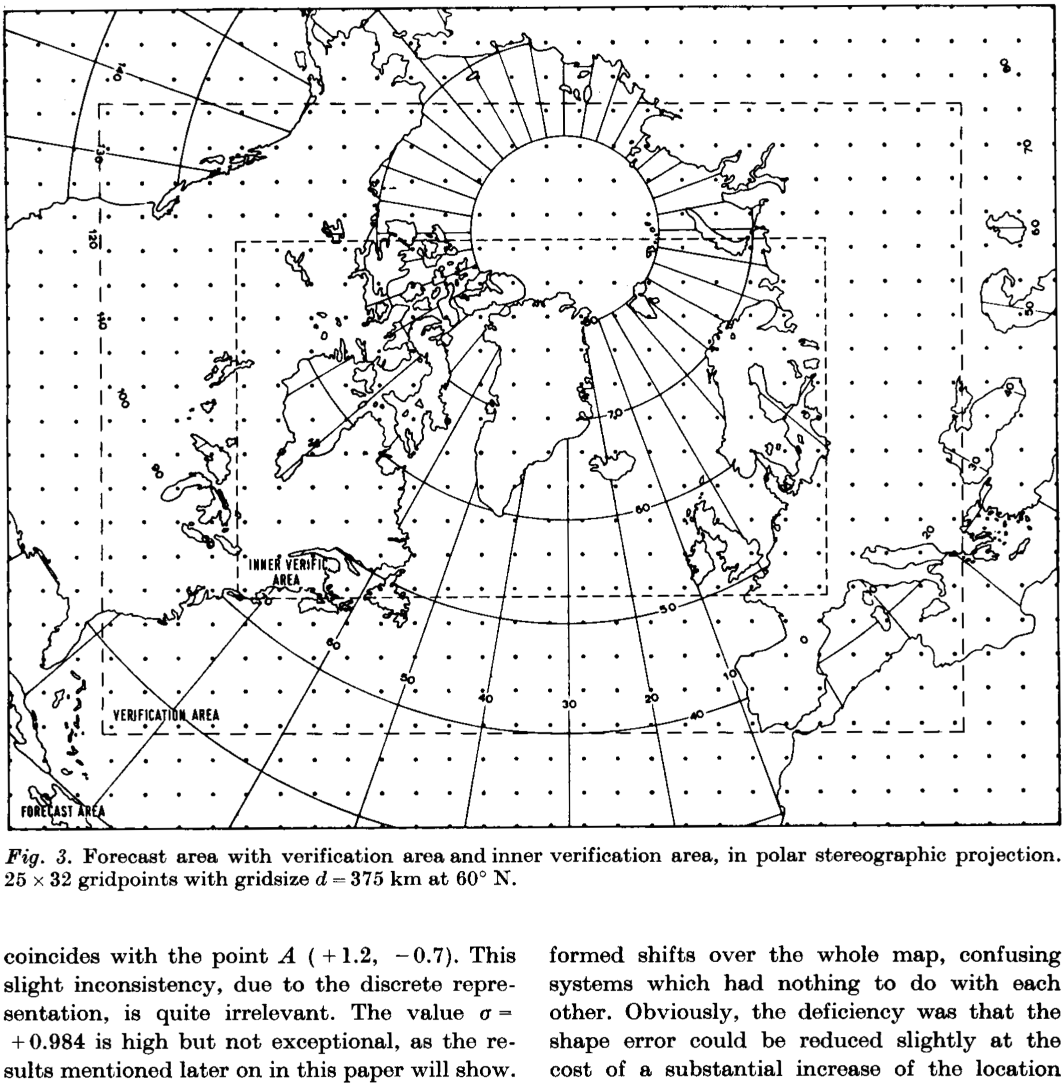


Fig. 3. Forecast area with verification area and inner verification area, in polar stereographic projection. 25×32 gridpoints with gridsize $d = 375$ km at 60° N.

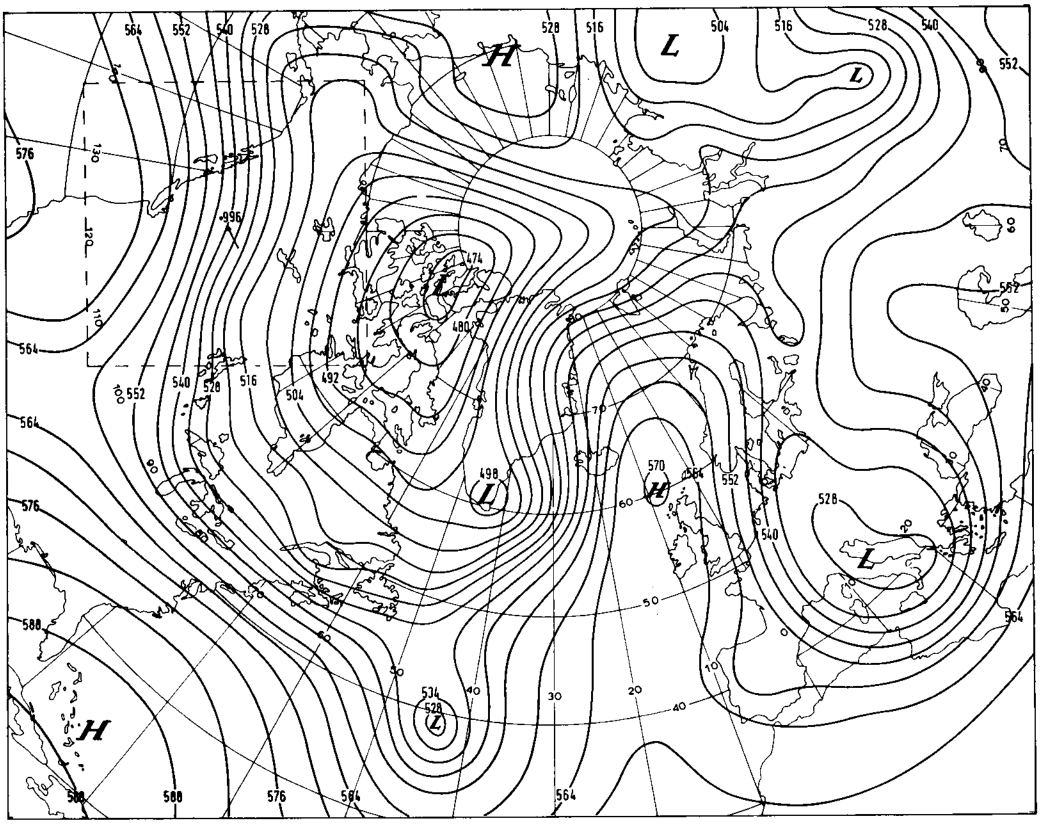
coincides with the point A $(+1.2, \ -0.7)$. This slight inconsistency, due to the discrete representation, is quite irrelevant. The value $\sigma = +0.984$ is high but not exceptional, as the results mentioned later on in this paper will show. The process should be repeated in the same way for all other possible reference systems lying within the area $|x|, |y| \leq 5$, but this is not carried out here.

The example demonstrates the separation of both effects, called I and II. If one is interested in the difference in height at the centre (III), this may easily be obtained from the data at gridpoints by using the interpolation procedure described in Appendix A.

In the first experiments a problem presented itself: sometimes an extremely large shift had to be made in order to obtain only a slight improvement of the s.c. Ambiguities even arose when the pattern recognition procedure performed shifts over the whole map, confusing systems which had nothing to do with each other. Obviously, the deficiency was that the shape error could be reduced slightly at the cost of a substantial increase of the location error. This difficulty could be overcome by imposing restrictions on the shifts to be allowed. A kind of damping mechanism, which implies that the idea $\sigma \overset{\mathrm{DEF}}{=\!=} \varrho_m$ is modified, provides for a balance between shape error and location error and will be explained in Appendix B. It should be borne in mind that this mechanism has been applied to obtain the following results.

3. Some preliminary results

For February 18, 1965, 0000 GMT, 24-hour and 48-hour barotropic forecasts were made of the geopotential heights of the 500 mbar pressure-surface. Fig. 3 shows the forecast area



a

Fig. 4 a. February 17, 1965, 0000 GMT: observed 500 mbar contourlines drawn for every 8 geopotential decameters. By dashed lines the system is delineated which shows the greatest geometrical resemblance to the system indicated in Fig. 4 *b*, with its location error vector and measure for the goodness of shape.

map in polar stereographic projection, covered by 25×32 gridpoints with a grid size of 375 km at 60° N. The map magnification was taken into account.

Each forecast was produced for two model versions: one without stabilization of the ultra-long waves (version *a*), the other in which the Cressman-term for correction of the retrogression of the ultra-long waves had been incorporated (version *b*). No terms accounting for the effects of topography and friction were included in both versions. The model used the complete balance equation. In Fig. 4 the actual situations of February 17 and 18, 1965, 0000 GMT are shown. Fig. 5 presents the 24-hour forecast and Fig. 6 the 48-hour forecast, each for version *b*.

Each forecast, and for comparison purposes also the situation of the 17th considered as a 24-hour "persistence" forecast, was verified according to the above outlined method with $m = n = 9$. Smaller values of m and n proved to make the method more sensitive to the small-scale phenomena, whereas greater values of m and n apparently stress the influence of the long waves. In choosing $m = n = 9$ the attention is focussed on an intermediate range of wave numbers, presumably coinciding with the scope of interest of the synoptic meteorologist. The results of such a small demonstration sample are not intended to serve as an ultimate judgement on whether inclusion of the Cressman-term would provide a significant improvement.

In the actual chart of February 18 (Fig. 4*b*) the boundary of one system to be verified is marked by dashed lines. In the forecast maps of Figs. 5 and 6 the interpolated system is shown which the computer found to have the greatest

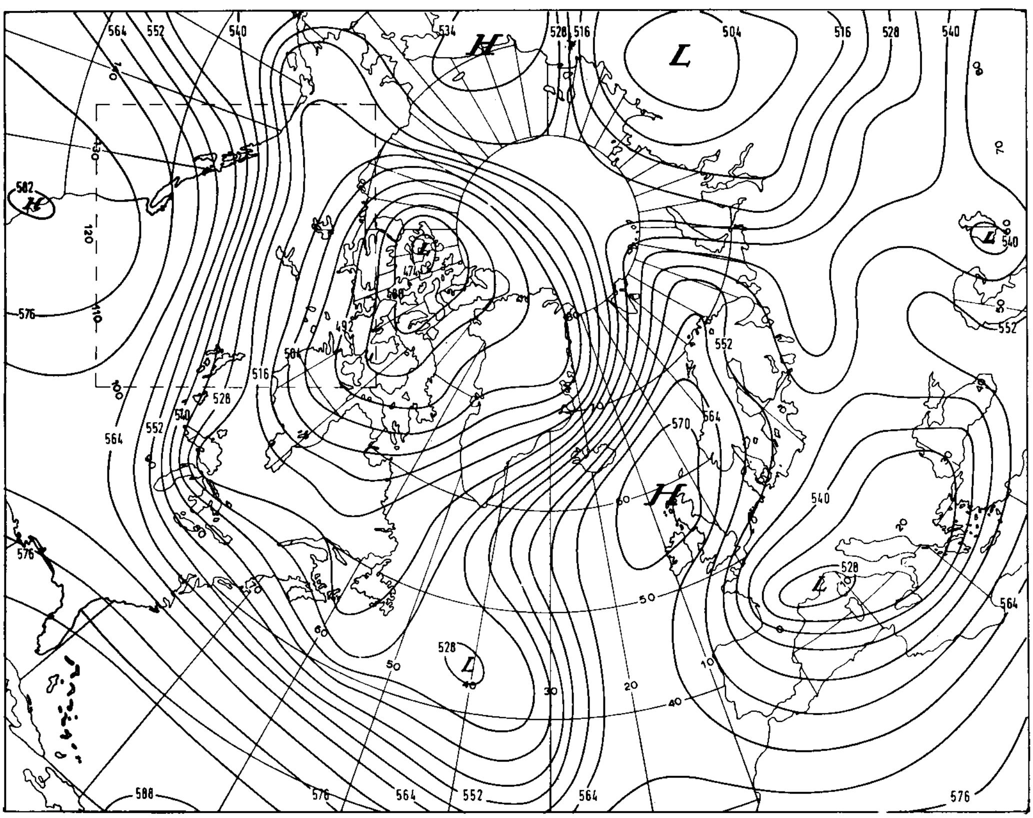


Fig. 4 b. The same as in Fig. 4 *a*, for February 18, with one of the systems to be verified ($m = n = 9$).

"resemblance" to the verifying system (apart from the slight reduction of the location error vector by the "damping mechanism" mentioned at the end of section 2). The arrows denoting ϵ, with the figures representing the corresponding values of σ, confirm subjective observation that in *that* region persistence is the best forecast and that the 24-hour forecasts are better than those for 48 hours, both in location and in shape. In the same way for all other verifying systems lying within the verification area, with centre-points that fall in the inner verification area (see Fig. 3), σ and ϵ were calculated. Figs. 7 and 8 represent the verifications referring to the respective maps shown in Figs. 5 and 6.

The daily verifications, stored on cards or tape, might be used for further processing and statistics. For example, time-averaging of the daily results over a long period may reveal certain types of *systematic* errors of the model, if any. To illustrate this idea, Figs. 9 and 10 contain verifications comparable to those of Figs. 7 and 8, but now obtained by averaging per gridpoint over the period 18 to 21 February. The features become already more regular even though, the period being still very short, the verifications only reveal the model errors associated with the type of circulation during these four days and not yet any *overall*-systematic errors of the model. Therefore, verifications should in practice be averaged over a month, a whole season or over all days with the same type of circulation during a year or so.

In Fig. 11 averaged verifications, some of which are shown in Figs. 9 and 10, have been summarized by means of frequency distributions. Figs. 11*a*, *c* and *e* show the distributions for σ, Figs. 11*b*, *d* and *f* for the length of the location error vector expressed in one-fourth of the gridsize d ($\lambda = 4|\epsilon|/d$). Figs. 11*a* and *b* refer to the 24-hour forecasts, Figs. 11*c* and *d* to the 48-hour forecasts, and Figs. 11*e* and *f*

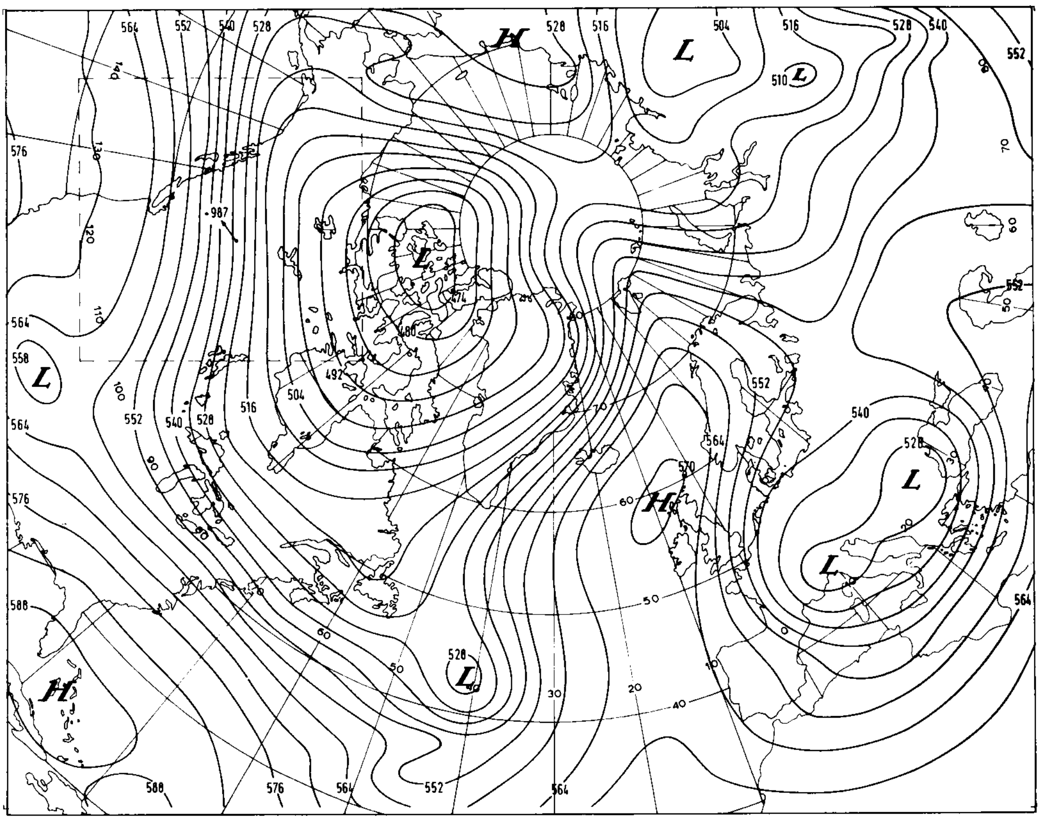


Fig. 5. 24-hours forecast with model version *b*, valid for February 18, 1965, 0000 GMT (cf. actual situation in Fig. 4 *b*). (Dashed-line square: see Fig. 4 *a*.)

to persistence. In Figs. 11*a*, *b*, *c* and *d* the white blocks refer to model-version *a* and the dark blocks to version *b*.

From the frequency distributions in Fig. 11 Table II has been derived in which the bar denotes an average calculated from frequencies per class.

Apparently, in this case version *b* has brought about an improvement in shape, but a deterioration occurs with respect to location in the 48-hour forecasts. The 48-hour forecasts only slightly overclass persistence concerning location; with version *a* the shape aspect is not better than with persistence.

Table II

	24-hr frcst		48-hr frcst		
	a	*b*	*a*	*b*	Persist.
$\bar{\sigma}$	0.964	0.970	0.868	0.918	0.886
$\bar{\lambda}$	0.895	0.870	2.060	2.255	2.565

4. Concluding remarks

In attempting to assess the degree to which two scalar fields resemble one another, two new indices are constructed. The author is aware of the shortcomings of all such indices, and in this particular case of the inherent correlation between the indices at neighbouring points.

In spite of this fact the method presented here might be helpful for the research worker in numerical weather prediction to give him more insight into the behaviour of his models with respect to the prediction of scalar fields. Varying the size of the "window", defined by m and n, could provide him with phase-infor-

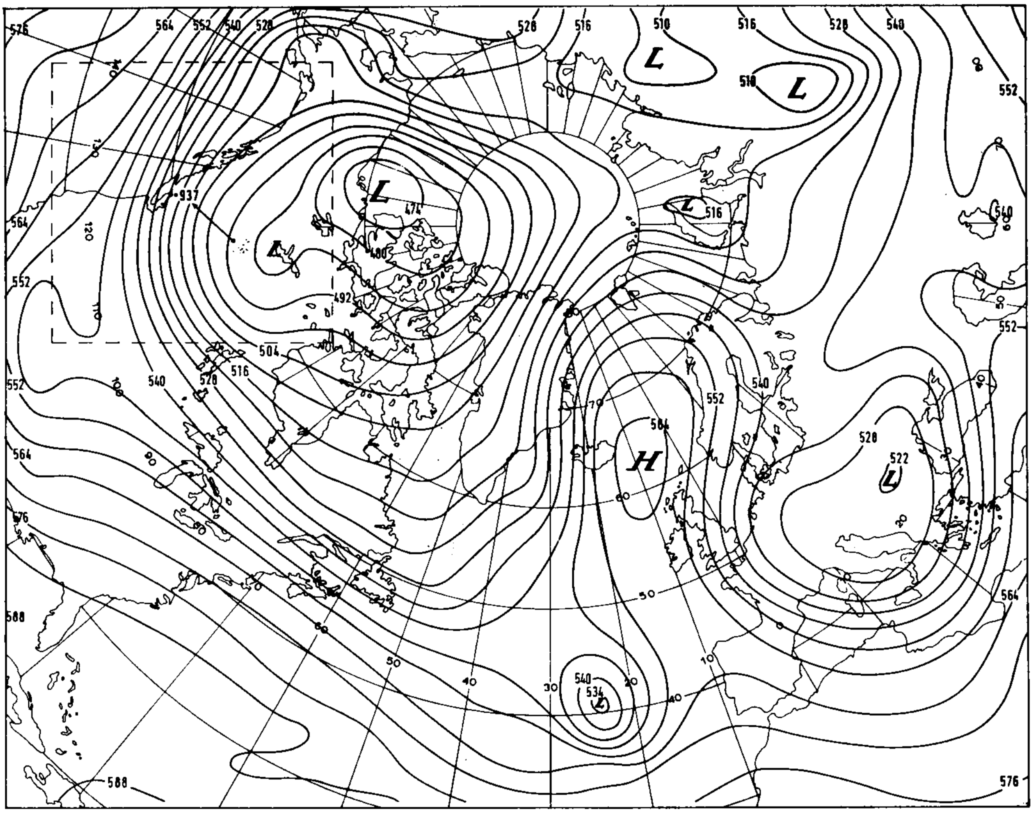


Fig. 6. The same as in Fig. 5, but 48-hour forecast.

mation for different ranges of wave numbers. An extension of the method would be to relate shape and location errors both to the earth's topography, to the temperature of the underlying surface or to vorticity—the parameter that is directly predicted.

Legend to figs. 7–10.
Classes for σ: A = 1.00–0.95, B = 0.95–0.90
C = 0.90–0.85, D = 0.85–0.80, E = 0.80–0.75.

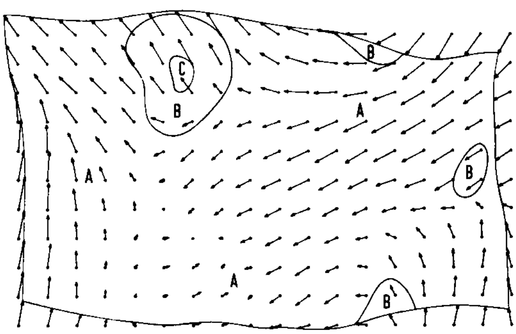


Fig. 7. Verification of the map shown in Fig. 5. The arrows denote location errors (see also Fig. 2), whereas isolines for σ are drawn. For further explanation see the legend to Figs. 7–10, which is shown aside.

Another possible aim, apart from checking the behaviour of numerical models by the research worker, would be the use of daily and averaged verification maps in the daily weather

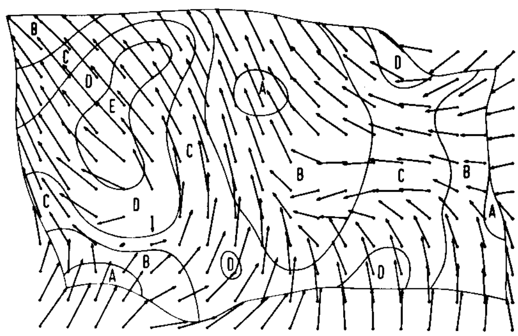


Fig. 8. Verification of the map shown in Fig. 6.

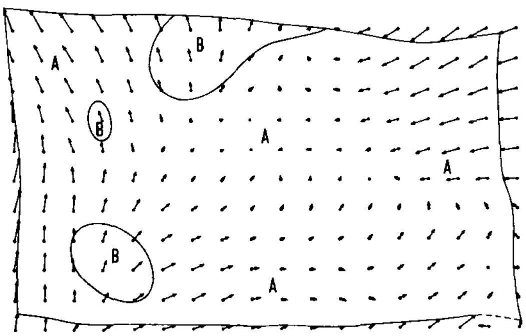


Fig. 9. Verification as in Fig. 7, but averaged over the period 18–21 February.

service. This might help synoptic meteorologists in making empirical corrections to the numerical prognoses, which at present is provided in a vague manner from practical experience. Ultimately, this adjusting procedure could be automated by making it a part of the forecasting routine in the computer.

5. Acknowledgements

The author wishes to express his gratitude to Prof. Dr. F. H. Schmidt for stimulating this work. Also Mr. H. C. Bijvoet is to be credited for his suggestions and many helpful discussions. Furthermore the author is indebted to Drs. W. J. A. Kuipers, Dr. C. Levert, Dr. C. J. E. Schuurmans and Drs. J. Wieringa for reading the manuscript carefully and critically.

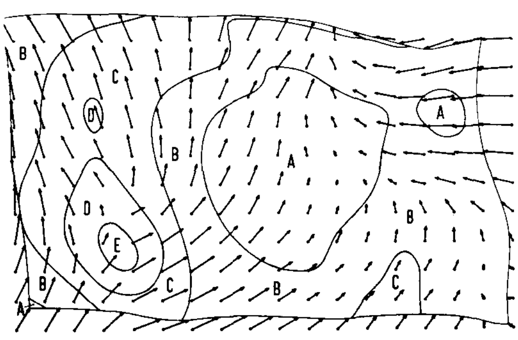


Fig. 10. Verification as in Fig. 8, but averaged over the period 18–21 February.

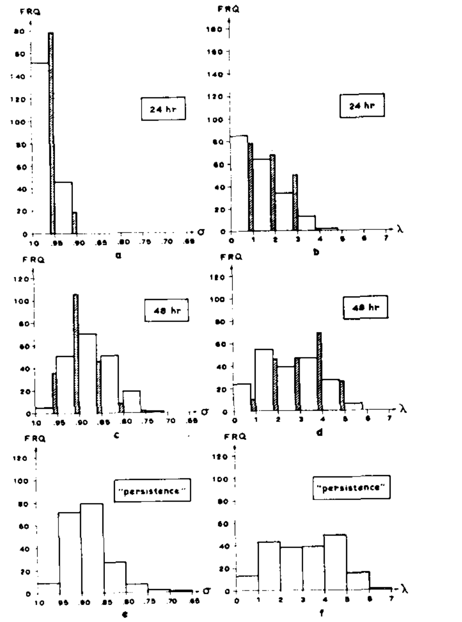


Fig. 11. Shape error (σ) and location error (λ) frequency distributions for 4-day averaged verifications. (a) and (b): 24-hour forecast. (c) and (d): 48-hour forecast. (e) and (f): persistence (24-hours). In (a), (b), (c) and (d) the white blocks refer to version a, the dark blocks to version b.

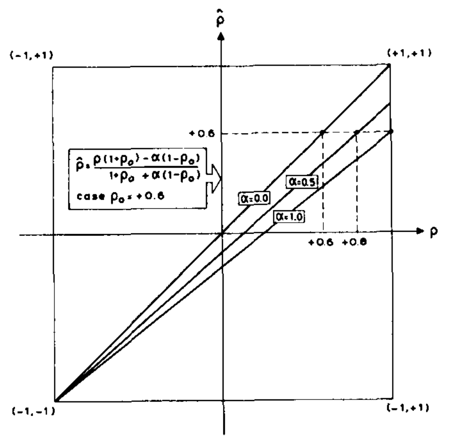


Fig. 12. (ϱ, $\hat{\varrho}$)-diagram, if $\varrho_0 = +0.6$, for $\alpha = 0.0$, $\alpha = 0.5$ and $\alpha = 1.0$, according to the formula (5).

REFERENCES

Bleeker, W. 1946. The verification of weather forecasts. Kon. Ned. Meteor. Inst. publ. 125: *Mededelingen en Verhandelingen (B), 1*, No. 2, 23 pp.

Brier, G. W. & Allen, R. A. 1951. Verification of weather forecasts. *Compendium of meteorology*, American Meteorological Society, Boston, Mass. pp. 841–848.

Holyoke, D. L. 1965. Verification of objective forecasts at the Central Analysis Office of the Meteorological Service of Canada, Canada—*Department of Transport—Meteorological Branch*, 24 pp.

Johnson, D. H. 1957. Forecast verification: A critical survey of the literature. *London, Air Ministry, Meteorological Research Committee*, M.R.P. 1056, S.C. 11/237, 40 pp.

Murphy, A. H. & Epstein, E. S. 1967. Verification of probabilistic predictions: a brief review. *Journal of Applied Meteorology 6*, pp. 748–755.

Appendix A

The interpolation procedure to obtain ϱ_m was set up as follows. Starting with the array of s.c.-values shown in Fig. 2, it is clear that a maximum will be found somewhere in the neighbourhood of Q, the gridpoint with the greatest value for the s.c. (+0.930). Therefore the second-degree polynomial

$$\varrho(x, y) = f \equiv a_0 + a_1 x + a_2 y + a_3 x^2 + a_4 xy + a_5 y^2 \tag{4}$$

with x and y as grid coordinates and the origin in Q, is fitted to the values at Q and at the four nearest surrounding gridpoints. To these five points one of the points K, L, M, N must be added in order to solve from (4) six equations with the six unknowns $a_0, a_1, \dots a_5$. It is convenient to take the point with the greatest s.c.-value, in this case L, so that the polynomial fits the values at the gridpoints labelled with a cross-symbol, as seen in Fig. 2. Then from the conditions $\partial f/\partial x = \partial f/\partial y = 0$, x_σ and y_σ are easily found. The arrangement of the program is such that not all values shown in Fig. 2 are calculated, but only as many values as needed to apply the interpolation procedure.

Now it will be clear that, if $\varrho_k = \varrho_0 + \alpha(1 - \varrho_0)$, $\hat{\varrho}_k$ must be put equal to ϱ_0. Therefore, if for each ϱ_0 a linear relationship is assumed between ϱ and $\hat{\varrho}$, in a $(\varrho, \hat{\varrho})$-diagram belonging to a given value of ϱ_0 the line should go through the point $\{\varrho_0 + \alpha(1 - \varrho_0), \varrho_0\}$. Passage of the line through $\{-1, -1\}$ is trivial. The equation of this line is

$$\hat{\varrho} = \frac{\varrho(1 + \varrho_0) - \alpha(1 - \varrho_0)}{1 + \varrho_0 + \alpha(1 - \varrho_0)} \tag{5}$$

Fig. 12 demonstrates, for $\varrho_0 = +0.60$, how $\hat{\varrho}$ depends on ϱ for some values of α.

Appendix B

Let ϱ_k denote a certain value of the s.c. associated with a shift over a distance equal to k times the gridsize d (the index k, which is not necessarily an integer quantity, will sometimes be omitted if the size of the shift is irrelevant). In order to suppress very large shifts, each value ϱ_k at gridpoints is weighted down to a new value, $\hat{\varrho}_k$, also ranging from -1 to $+1$. The greater k, the more ϱ_k is pressed down.

Now the location error will not be coupled any longer with the position of ϱ_m, the interpolated maximum of the field of ϱ-values, but with the position of the interpolated maximum of the field of $\hat{\varrho}$-values, $\hat{\varrho}_m$. The measure for the goodness of shape remains associated with the unmodified ϱ-values, though not with their maximum value ϱ_m, but with the ϱ-value corresponding with the *position* of $\hat{\varrho}_m$. In other terms: $\sigma \stackrel{\text{DEF}}{=} \varrho(x_{\hat{\varrho}_m}, y_{\hat{\varrho}_m})$.

This artifice balances enlargement of the location error against improvement in shape. This process depends on the nature of the damping, which must be specified as a function of k. Unfortunately, this can only be done in a subjective way.

For example, if $\varrho_0 = +0.60$, it could be required that ϱ_3 ought to attain the value $+0.80$ at least so as to justify longer shifts. In relation to the above this means that, if $\varrho_3 \leq +0.80$, $\hat{\varrho}_3$ may not exceed the value $+0.60$. In this case, at least 50% of the greatest possible improvement (Δ) of the s.c. with respect to ϱ_0, where $\Delta = 1 - \varrho_0 = +0.40$, must be bridged to allow for a location error vector with length $\geq 3d$.

Generally it might be required that a shift over the distance $k \cdot d$ must result in a s.c. improvement of at least a fraction α of the possible improvement Δ, where $\alpha = \alpha(k)$. Tentatively α

has been chosen in the following completely subjective way

$$\alpha = 1 - \exp(-k^3/40).$$

From this it follows that for shifts over $2d$, $3d$ and $4d$ the improvement of the s.c. with respect to ϱ_0 must be at least about 18, 49 and 80 per cent of Δ, respectively, to justify any further shifts.

This approach implies that an improvement from $+0.60$ to $+0.80$ is equally valued with an improvement from $+0.80$ to $+0.90$ or from $+0.20$ to $+0.60$ etc. Possibly another approach or the use of a different function $\alpha(k)$ may prove more reasonable, though the author believes such modifications not to be essential for the method.

НОВЫЙ МЕТОД ОЦЕНКИ ДЕТЕРМИНИСТИЧЕСКИХ ПРОГНОЗОВ СКАЛЯРНЫХ МЕТЕОРОЛОГИЧЕСКИХ ПОЛЕЙ

При численном прогнозе погоды большинство статистических характеристик ошибок, относящихся к прогнозу давления, определяются суммарным действием ошибок «формы» и ошибок местоположения. Желательно разделить эти эффекты. Предлагается метод, дающий возможность сделать это для краткосрочных прогнозов, если пренебречь возможными вращательными эффектами. Схема еще не была испытана в оперативной практике, но в заметке обсуждаются результаты ее применения в некоторых примерах. В работе не затрагиваются вопросы вероятностного прогноза и экономические аспекты оценок с точки зрения потребителей, так как наша цель — это только научные вопросы оценки поведения численных моделей.