Data Assimilation by Statistical Interpolation of Forecast Error Fields

IAN D. RUTHERFORD

Atmospheric Environment Service, Dorval, Quebec, Canada

(Manuscript received 7 September 1971, in revised form 10 March 1972)

ABSTRACT

An operational method of combining observations with short-period forecasts of the same quantities is described. The method amounts to performing an interpolation in the field of apparent forecast errors, as defined by the available observations, by a linear least-squares technique similar to Gandin's. Statistics on the errors of observation and prediction are required. The spatial autocorrelations of the forecast error depend on the particular forecast model employed, the forecast period, etc. Those statistics necessary for the assimilation of synoptic data with four-level baroclinic 12-hr forecasts have been derived. They appear to be approximately homogeneous and isotropic and to vary slowly with season. Asynoptic data can be handled by allowing for variations with forecast period. Some practical problems in the determination of empirical statistics are discussed.

1. Introduction

Most operational objective analysis systems employ a short-period forecast as a guess or background field. In the sense that the final analysis contains contributions from both the current data and from the forecasts (based on past data) these are methods of data assimilation. However, the emphasis has usually been on "fitting" the current data as closely as possible and the forecasts have been regarded as only a necessary evil, needed to fill in the holes in the observing network. In fact, short-range forecasts have an error level not much larger than the error of a single observation and they ought to be weighted accordingly. In addition, the influence region of a single observation has usually been rather arbitrarily prescribed.

The purpose of this paper is to explain a method of assimilating data with short-range forecasts which ensures an optimum relative weighting and a statistically best estimate of the true field values. The method is similar to Gandin's (1963) method of "optimum" interpolation applied to the field of observed-minus-forecast differences rather than to the observations themselves. Allowance is made for nonhomogeneity of the mean and variance of these residuals. Some problems in determining the necessary autocorrelation functions of the prediction error and their smoothing and extrapolation to short lags are discussed. A method of obtaining these and other statistics as a by-product of the data-checking and interpolation procedures is suggested.

2. Assimilation equations

The problem of data assimilation may be regarded as the correction of a field of predicted values of some quantity ψ in the light of observations ψ_j^0 scattered irregularly in space and time, to form an estimate of the most likely true field, taking account of the errors of prediction and observation. Let us consider for the moment that the predictions and observations all pertain to the same time and concern ourselves only with the spatial weighting problem. This is "intermittent" data assimilation in the terminology of the preliminary GARP Working Group Report on Four-Dimensional Data Assimilation.

Let us correct the predicted ψ_k^p at field location k using a linear combination of the apparent forecast errors at the set of observing stations i, $i=1, 2, \cdots, N$. That is, let us define an estimate, ψ_k^e, of the true value, ψ_k^t, at k by

$$\psi_k^e=\psi_k^p+\sum_{j=1}^{N}\alpha_{kj}(\psi_j^0-\psi_j^p), \tag{1}$$

subject to the least-squares minimization condition that

$$E_k^2\equiv\overline{(\psi_k^t-\psi_k^e)^2} \tag{2}$$

be a minimum, where E_k^2 is the mean square error of estimation (interpolation) and the overbar indicates an ensemble average over many realizations of the assimilation procedure. The normal equations, whose solutions are the weight coefficients α_{kj}, are then easily written as

$$\sum_{j=1}^{N}\alpha_{kj}\overline{(\psi_i^o-\psi_i^p)(\psi_j^o-\psi_j^p)}=\overline{(\psi_k^t-\psi_k^p)(\psi_i^o-\psi_i^p)},$$
$$i=1, 2, \cdots, N. \tag{3}$$

Introducing $\epsilon_i^o=\psi_i^t-\psi_i^o$ and $\epsilon_i^p=\psi_i^t-\psi_i^p$, and assum-

ing that the observational errors are random, i.e., that

$$\overline{\epsilon_i{}^o\epsilon_j{}^p}=0 \quad \text{for all } i,j, \tag{4}$$

$$\overline{\epsilon_i{}^o\epsilon_j{}^o}=\delta_{ij}\sigma_o{}^2, \tag{5}$$

where δ_{ij} is the Kronecker delta and $\sigma_o{}^2$ the (constant) observational error variance, then the normal equations can be written

$$\sum_{j=1}^{N}(\overline{\epsilon_i{}^p\epsilon_j{}^p}+\delta_{ij}\sigma_o{}^2)\alpha_{kj}=\overline{\epsilon_k{}^p\epsilon_i{}^p}, \quad i=1, 2, \cdots, N. \tag{6}$$

Assumptions (4) and (5) state that the observational errors are uncorrelated with each other or with the prediction errors and have the same variance everywhere. The prediction error covariances can be obtained empirically, as will be discussed later.

It has been implicitly assumed that

$$(\psi_i{}^o-\psi_i{}^p)=0. \tag{7}$$

We also made the further assumption that the total, i.e., prediction plus observational, error variance

$$\sigma^2=\overline{(\psi_i{}^o-\psi_i{}^p)^2}=\text{a constant}. \tag{8}$$

The above procedure is equivalent to Gandin's (1963) method of optimum interpolation applied to the field of apparent forecast errors or residuals $\psi^o-\psi^p$ rather than to ψ^o alone. It was developed independently by Eddy (1964, 1967) and Kruger (1964, *et seq.*), and has been used operationally at the Central Analysis Office in Montreal with certain simplifications, for a number of years. It amounts to estimating the forecast error at the grid points k by ordinary multiple linear regression on the apparent forecast errors at surrounding observing stations. It differs from the Cressman (1959) method, employed in the United States and elsewhere, mainly in that the weights are determined statistically and are dependent on the number and distribution of reports rather than on distance from the grid point alone (Kruger, 1969). It reduces to something like the Cressman method if the left-hand matrix of the normal equations is replaced by the unit matrix.

Assumptions (7) and (8) regarding spatial uniformity of the mean and variance of the observed-minus-forecast differences may be relaxed somewhat simply by standardizing the differences by subtraction of their local mean and division by the local standard deviation, σ_j, assuming that these are available. In that case, the assimilation equation (1) becomes

$$\psi_k{}^e=\psi_k{}^p+\overline{(\psi_k{}^o-\psi_k{}^p)}+\sigma_k\sum_{j=1}^{N}\alpha_{kj}\left(\frac{\psi_j{}^o-\psi_j{}^p-\overline{(\psi_j{}^o-\psi_j{}^p)}}{\sigma_j}\right), \tag{9}$$

and the normal equations (6) become

$$\sum_{j=1}^{N}\left(\mu_{ij}+\delta_{ij}\frac{\sigma_o{}^2}{\sigma_i\sigma_j}\right)\alpha_{kj}=\mu_{ki}, \quad i=1, 2, \cdots, N, \tag{10}$$

where

$$\mu_{ij}=\frac{\overline{(\epsilon_i{}^p-\bar{\epsilon}_i{}^p)(\epsilon_j{}^p-\bar{\epsilon}_j{}^p)}}{\sigma_i\sigma_j} \tag{11}$$

are the normalized prediction error covariances. Note that

$$\mu_{ii}=\frac{\sigma_{p,i}{}^2}{\sigma_i{}^2}=\frac{\sigma_{p,i}{}^2}{\sigma_{p,i}{}^2+\sigma_o{}^2}. \tag{12}$$

Because of assumptions (4) and (5) and for $i\neq j$,

$$\mu_{ij}=\frac{\overline{[(\psi_i{}^o-\psi_i{}^p)-\overline{(\psi_i{}^o-\psi_i{}^p)}][(\psi_j{}^o-\psi_j{}^p)-\overline{(\psi_j{}^p-\psi_j{}^p)}]}}{\sigma_i\sigma_j}, \tag{13}$$

the correlation coefficient between apparent forecast errors at i and j.

A single isolated report exactly at a grid point would be assimilated with a weighting factor (12) and the corrected value of the field would be

$$\psi_k{}^e=\frac{\sigma_{p,k}{}^2\sigma_o{}^2}{\sigma_{p,k}{}^2+\sigma_o{}^2}\left(\frac{\psi_k{}^p}{\sigma_{p,k}{}^2}+\frac{\psi_k{}^o}{\sigma_o{}^2}\right)+\frac{\sigma_o{}^2}{\sigma_{p,k}{}^2+\sigma_o{}^2}\overline{(\psi_k{}^o-\psi_k{}^p)}. \tag{14}$$

The observation and the prediction are each weighted inversely as their respective gross error variances.

3. Interpolation error

The interpolation error, or mean square error of estimation, as defined in (2) using the assimilation equation (9) with coefficients obtained from the solution of (10) can be easily shown to be

$$E_k{}^2=\sigma_{p,k}{}^2-\sigma_k{}^2\sum_{j=1}^{N}\alpha_{kj}\mu_{kj}. \tag{15}$$

If no observations were available the best estimate $\psi_k{}^e$ would be simply $\psi_k{}^p+(\psi_k{}^o-\psi_k{}^p)$ and it would have error $E_k{}^2=\sigma_{p,k}{}^2$. For the example just given of a single isolated report the error at the report location is

$$E_k{}^2=\left(\frac{\sigma_{p,k}{}^2}{\sigma_{p,k}{}^2+\sigma_o{}^2}\right)\sigma_o{}^2.$$

This interpolation error is less than $\sigma_o{}^2$, the error which corresponds to simple replacement of the predicted value by the observed value and, if $\sigma_{p,k}{}^2$ is not much larger than $\sigma_o{}^2$, $E_k{}^2$ is considerably less than $\sigma_o{}^2$.

4. Determination of correlation coefficients

In principle, Eq. (9) should be applied to all observations in three-dimensional space. In practice, it is necessary to limit consideration to that volume for which the correlations are large, which usually means the immediate neighborhood of the field point considered. The assimilation is four-dimensional in the sense that predictions, which depend on previous data, are combined with current data. Here, however, consideration will be given only to data on the same constant pressure surface. This is three-dimensional assimilation with the missing dimension being the vertical rather than time. In addition, the time variable will be highly discretized, to multiples of 12 hr.

The correlation coefficient matrix $\{\mu_{ij}\}$ should be determined uniquely for each combination of observing locations and for each assimilation cycle. Each of the elements, if known at one time, could be forecast for the next cycle. In practice this is difficult if not impossible with nonlinear prediction equations. The matrix may be determined empirically, however, by averaging products of prediction errors for an ensemble of cases. If they turn out to be homogeneous, as is often assumed from the outset, then each of the μ_{ij} can be determined simply from the geometry of the situation. If, in addition, they are isotropic, the μ_{ij} depend simply on separation. Homogeneity and isotropy were assumed by Gandin (1963), Eddy (1967) and Kruger (1969) for most fields. Petersen and Truske (1969) assumed only homogeneity. In operational practice one is forced to assume homogeneity and if not isotropy then some simple behavior as a function of azimuth. Even then, the resulting distance-dependent functions can be expected to depend as well on at least the particular forecast model employed, on the period of the forecast, the season of the year, etc.

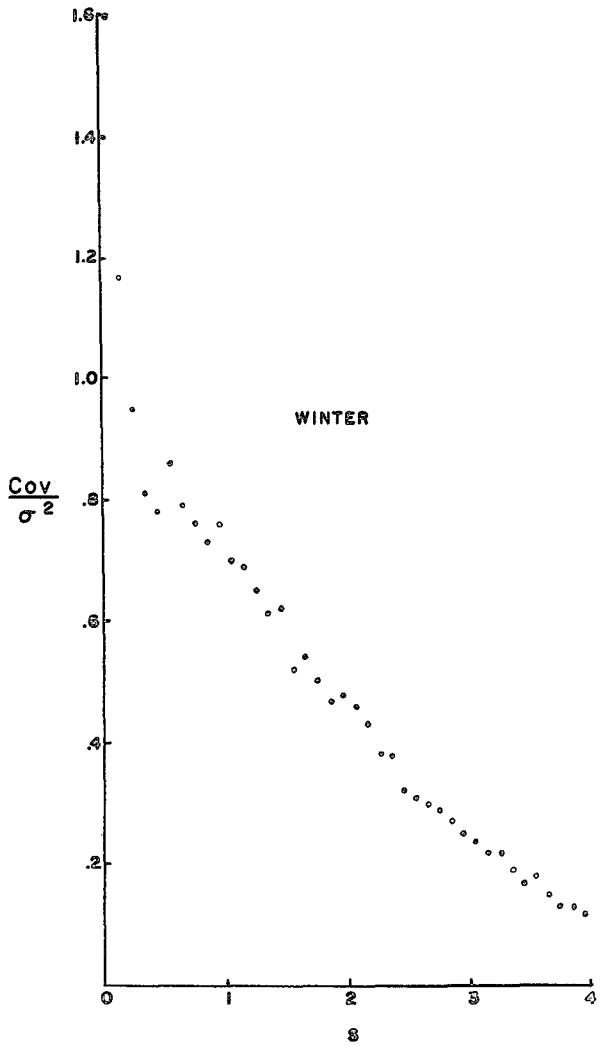

FIG. 1. Raw normalized autocovariance of residuals vs separation (in units of $d=381$ km) for 12-hr forecasts of 1000-mb height for winter 1969.

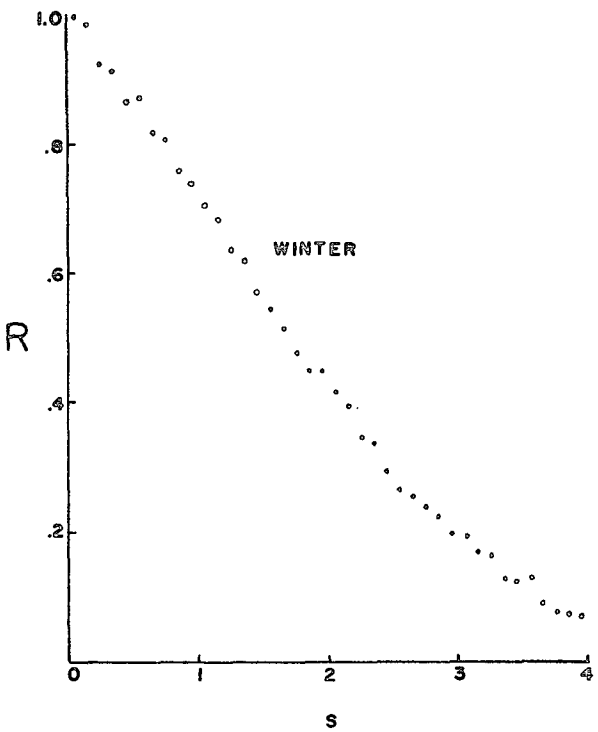

FIG. 2. Raw autocorrelation coefficients of residuals vs separation (in units of $d=381$ km) for 12-hr forecasts of 1000-mb height for winter 1969.

Covariances and correlations were determined empirically for the differences between observed geopotential heights and 12-hr forecasts for the four levels of the CAO operational baroclinic model. The data consisted of all the observations and corresponding interpolated forecasts for the period September 1969 to July 1970, broken down into four seasonal samples, each of three months, except summer which had only two because of technical data difficulties. Data at 1000 mb was augmented by values computed from surface

synoptic observations. Statistics were calculated for all possible combinations of reports separated by less than four standard grid lengths (1524 km) and these were averaged for each of 40 separation intervals of width 38.1 km. That is, homogeneity and isotropy were assumed. The number of station pairs is roughly proportional to the mean separation for the band and there are none at all for the two shortest separations, except at 1000 mb. Smoothing and extrapolation of these raw curves will be discussed in the next section. It soon became apparent that the variances were not horizontally uniform and that, therefore, the auto-correlations should satisfy the assumption of homogeneity more closely than the autocovariances. This was borne out by the appearance of somewhat less scatter in the former, as shown in Figs. 1 and 2.

5. Spectral smoothing and extrapolation of raw autocorrelations to short lags

A two-dimensional isotropic correlation function can be written as

$$\mu_{ij}=R(s),$$

where

$$s=[(x_i-x_j)^2+(y_i-y_j)^2]^{\frac{1}{2}}$$

is the separation. The power spectrum $S(k)$ can be calculated by Fourier transformation of $R(s)$, which, under isotropy, becomes a Hankel transform (see e.g., Gandin, 1963). That is, $R(s)$ and $S(k)$ are related by

$$R(s)=\int_0^\infty S(k)J_0(ks)kdk, \tag{16}$$

$$S(k)=\int_0^\infty R(s)J_0(ks)sds, \tag{17}$$

where J_0 is a Bessel function of the first kind and order zero and k is the magnitude of the (two-dimensional) wavenumber $2\pi/L$, where L is wavelength. A factor $1/(2\pi)$ has been absorbed into S. This is for continuous variables s and k and infinite range. For discrete discontinuous variables and finite range the integrals become quadrature sums, i.e.,

$$R(s_j)=\sum_{i=1}^{N} S(k_i)J_0(k_is_j), \quad j=1, 2, \cdots, N, \tag{18}$$

where N is the number of separation intervals for which R is tabulated. A factor k_i has been absorbed into the definition of the coefficients $S(k_i)$. The wavenumber scale $k_1, k_2, \cdots, k_N$ is determined by the requirement that the functions $J_0(k_is)$ be orthogonal to each other and to a constant over the domain $[0,s_N]$ and hence are given by the roots of $J_1(ks_N)=0$. The $S(k_i)$ are obtained by least-squares fitting of (18) to the available $R(s_j)$ and are therefore found as solutions to

$$\sum_{i=1}^{N}\sum_{l=1}^{N} w_lJ_0(k_is_l)J_0(k_js_l)S(k_i)=\sum_{l=1} w_lR(s_l)J_0(k_js_l),$$

$$j=1, 2, \cdots, N, \tag{19}$$

where the w_l are quadrature weights. Now if these are taken as $s_l\Delta s$ where s_l is the mean separation for the interval and Δs is the interval width then the left-hand matrix of (19) is nearly diagonal (the off-diagonal elements will not be exactly zero because the J_0's are not exactly orthogonal in the guadrature domain) and (19) reduces to a quadrature analog of (17). Fitting was done with both these weights and with $w_l=n_l$, the number of report pairs contributing to the raw correlation in the lth separation interval. Very little difference was noted since n_l is roughly proportional to s_l in any case.

Theoretically, it should be possible to fit the N raw estimates exactly using N coefficients. However, for purposes of objective analysis it would seem to be undesirable to include power from scales which are unrepresentable on the finite-difference grid. Truncation of the spectrum at some $k_{\max}\leqslant\pi/d$ where d is the grid length gives a natural way of smoothing the raw curve. The transform of the truncated spectrum can then be extrapolated to zero separation. In any case it is undesirable to include any component with negative power (Gandin, 1963; Eddy, 1967). Examples of truncated line spectra obtained in this manner are shown in Figs. 3 and 4.

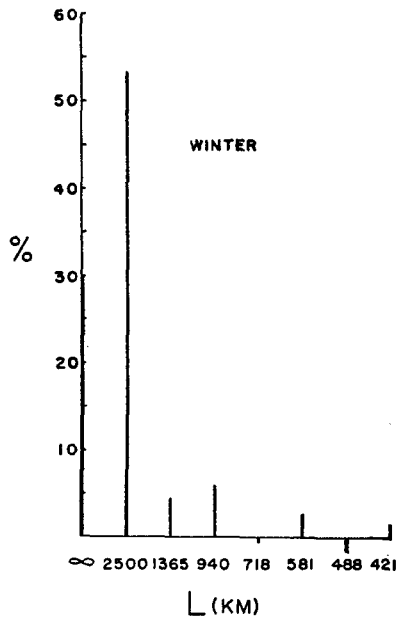

FIG. 3. Truncated line spectrum: percentage of variance vs wavelength (km) for 12-hr forecasts of 1000-mb height for winter 1969.

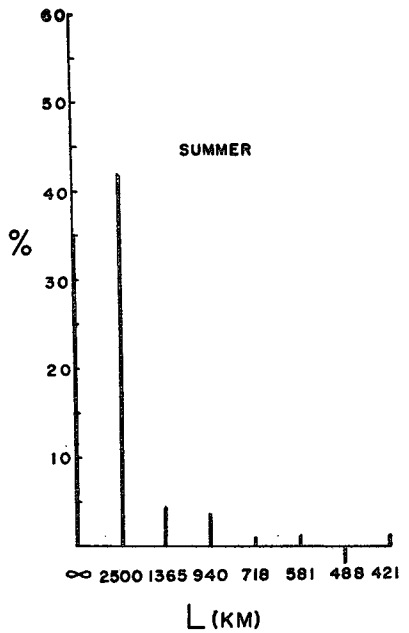

FIG. 4. Same as Fig. 3 except for summer 1970.

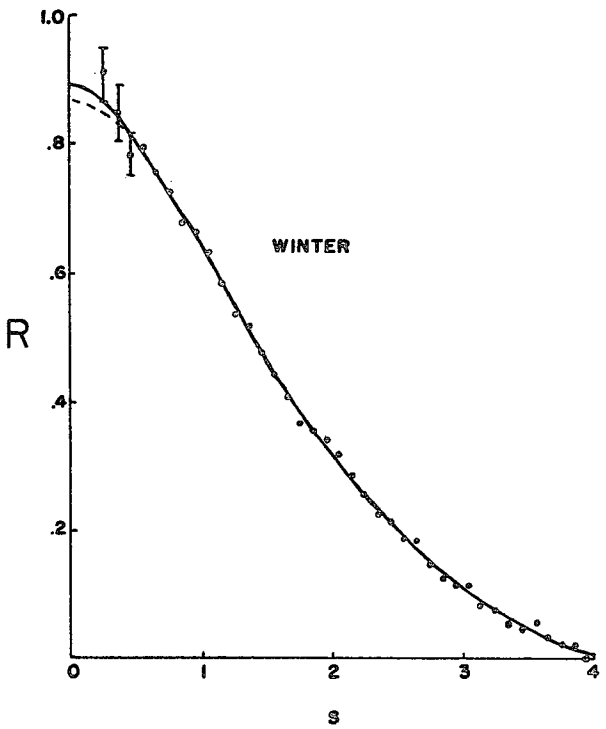

FIG. 5. Raw autocorrelation coefficients and final smoothed and extrapolated autocorrelation functions for 12-hr forecasts of 500-mb height for winter 1969. The dashed curved includes all power from wavelengths $>2d$ while the full curve includes in addition the power from wavelengths between $2d$ and d. The error bars correspond to $\pm\sigma_R$ where σ_R is the standard deviation of R assuming $N/3$ degrees of freedom.

Very little power is found in scales $\lesssim 800$ km and that is frequently negative. Negative power can appear because of the fact that we are simply fitting a set of functions to raw data with no restriction on the sign of the coefficients. The negative coefficients could be set to zero and the others adjusted to maintain the same total power but in fact we are not too interested in the short wave end of the spectrum anyway.

Fig. 5 shows a final smoothed and extrapolated autocorrelation function for two different truncations. One includes power from scales larger than $2d$ and the other includes scales down to d. The only difference between the two curves visible in the diagram is at small separations. The first would be appropriate for objective analysis or data assimilation on the normal grid while the second could be used with half the normal grid spacing. The extrapolated correlation at zero separation is equal to the ratio:

$$\left(\frac{\text{prediction error variance in large scales}}{\text{total prediction error variance} + \text{observation error variance}}\right).$$

Since the denominator is known, a separation into large-scale prediction error and the sum of small-scale prediction error and observational error can be made. Table 1 gives this breakdown for the four levels and for summer 1970 and winter 1969. In the discussion which follows the expression "observational error" should be understood to include sub-grid-scale prediction error.

Table 1 and Figs. 6 and 7 show that the observational error variance increases with height more rapidly than the 12-hr prediction error variance. Therefore, at high levels the data gets less weight and the forecast proportionately more. There is more smoothing of the data. The observational error remains nearly constant with season whereas the prediction error is much larger in winter than in summer, as is well known; the data get more weight in winter. There appears to be a change in the spectral distribution of the prediction error with

TABLE 1. Standard deviation of large-scale prediction error (σ_p) and of observation error plus small-scale prediction error (σ_o), for the four levels of the CAO filtered baroclinic model (12-hr forecasts: units, meters).

Season		Pressure (mb) 1000	850	500	200
Winter	σ_p	44.1	37.4	43.7	49.0
	σ_o	11.6	10.1	17.3	31.5
Summer	σ_p	29.4	21.9	28.7	39.6
	σ_o	12.3	9.2	17.1	30.1

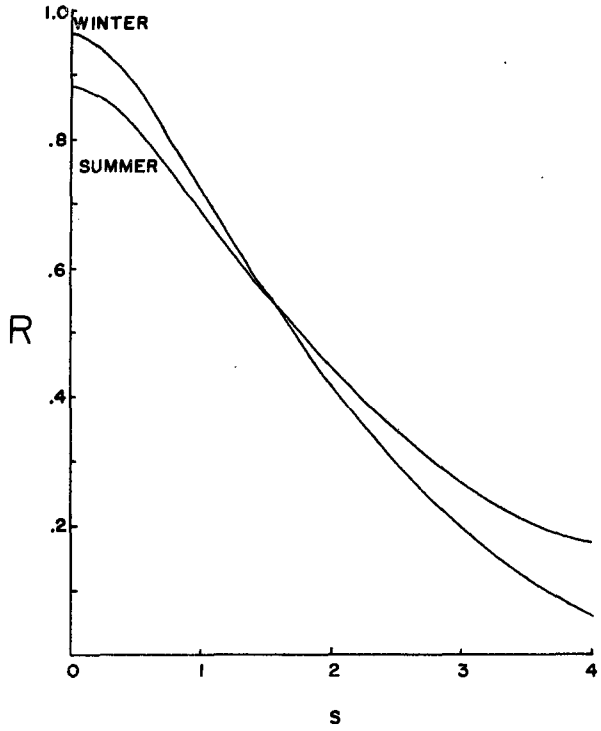

FIG. 6. A comparison of smoothed autocorrelation functions for two seasons for 1000-mb residuals.

season such that the autocorrelation drops off less rapidly with distance in summer; a larger radius of influence is indicated.

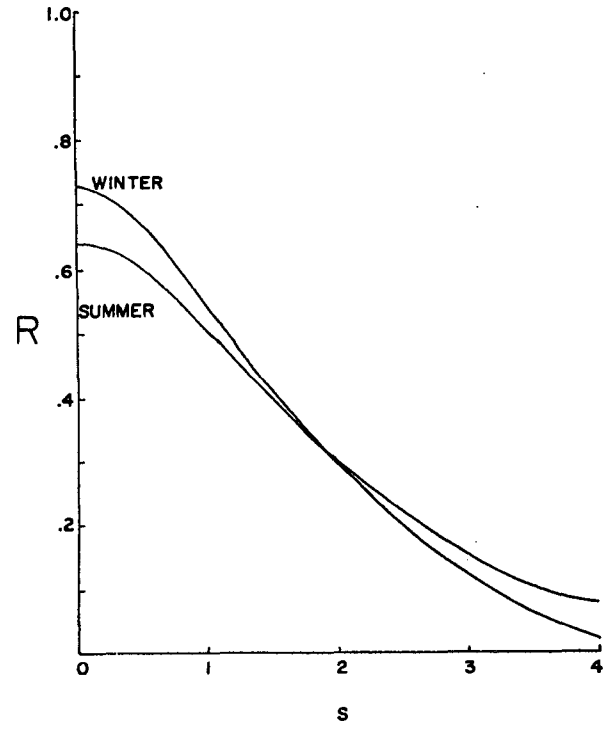

FIG. 7. Same as Fig. 6 except for 200-mb residuals.

6. Data checking and accumulation of statistics

The logical way to check a piece of data before assimilation would seem to be to make a preliminary interpolation at the location of the report excluding the report itself. Again, the quantity interpolated is the observed-minus-forecast difference. The interpolated difference is then compared with the reported one and the report accepted or rejected according to whether the discrepancy is smaller or larger than some suitable multiple of the interpolation error, as given by (15). This procedure automatically varies the tolerance according to the density and distribution of neighboring data. It amounts to comparing each report to an analysis done without that report. It is not too time-consuming because the number of reports is usually much less than the number of grid points (about 500 compared to 3000, typically). It has the added advantage that in performing the preliminary interpolation each data pair within the radius of influence is examined and it is a simple matter to accumulate products of residuals sorted according to their separation and thus accumulate the covariance statistics as a by-product. Sums can be accumulated for any desired period but something of the order of a month has been found to be necessary for reasonable stability.

A weighting scheme for partial sums like

$$\bar{S}^n=\nu S_n+(1-\nu)\bar{S}^{n-1}, \tag{20}$$

where $\overline{(\ )}^n$ indicates a weighted running sum after n assimilation cycles and S_n is the partial sum for the nth cycle, gives weight ν to the nth contribution, $\nu(1-\nu)$ to the $(n-1)$st, and so on. This gives automatic updating of the running mean, albeit with a certain time lag. In addition to the sums of cross products, fields of the mean and variance of the residuals are required and these can be easily accumulated once the analysis is complete.

7. Conclusion

The proposed assimilation scheme weights the forecasts and new data according to gross error statistics. It could be extended to account for correlation in the vertical and to cross correlation between related variables. Asynoptic data could be handled by allowing for the variation of prediction error with forecast period. Some information already exists on the variation of the spectral distribution with forecast period. Some sort of self-updating of statistics such as that proposed would seem to be necessary since the calculation and tabulation of these for all the variables and models in a typical NWP operation would be an impossibly large task, and it would have to be redone each time a model change was introduced.

REFERENCES

Cressman, G. P., 1959: An operational objective analysis system. *Mon. Wea. Rev.*, **87**, 367–374.

Eddy, A., 1964: The objective analysis of horizontal wind divergence fields. *Quart. J. Roy. Meteor. Soc.*, **90**, 424–440.

——, 1967: The statistical objective analysis of scalar data fields. *J. Appl. Meteor.*, 4, 597–609.

Gandin, L. S., 1963: *Objective Analysis of Meteorological Fields*. Leningrad, Gidromet.; Jerusalem, Israel Program for Scientific Translations, 1965.

Kruger, H. B., 1964: A statistical-dynamical objective analysis scheme. *Canadian Meteorological Memoirs*, No. 18, Meteorological Branch, Department of Transport, Toronto; also supplements No. 1 (1967, *CMM No. 23*) and No. 2 (1969, *CMM No. 27*).

——, 1969: General and special approaches to the problem of objective analysis of meteorological variables. *Quart. J. Roy. Meteor. Soc.*, **95**, 21–39.

Peterson, D. P., and T. N. Truske, 1969: A study of objective analysis techniques for meteorological fields. Final Rept EE-163(69) DC-104, Bureau of Engineering Research, University of New Mexico, Albuquerque.