U.S. DEPARTMENT OF COMMERCE
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
NATIONAL WEATHER SERVICE
NATIONAL METEOROLOGICAL CENTER

OFFICE NOTE 216

Evolution and Present Status of Objective Analysis/Assimilation at NMC

Ronald D. McPherson
Development Division

JULY 1980

I. Introduction

At the beginning of the numerical weather prediction era, initial conditions for the early numerical predictions were obtained from manually-analyzed meteorological charts. Values of the analyzed parameters were laboriously interpolated to the intersections of the grid point lattice upon which the finite-difference model equations were to be solved, and then manually entered on punched cards. It was quickly evident that automation of this process would be essential if operational numerical weather prediction was to be successful.

The Joint Numerical Weather Prediction (JNWP) Unit, supported jointly by the U.S. Air Force, Navy, and Weather Bureau, was established in July 1954 and first began making scheduled numerical predictions in April 1955 (Staff, JNWPU: 1957). Subjective analyses for these predictions were provided by the National Weather Analysis Center. By July 1955, experiments were in progress with an automated procedure, or "objective analysis." An operational version was implemented on 10 October 1955. Thus only 6 months elapsed between the first scheduled prediction and the first operational objective analysis.

A review of the papers and other documents describing the first few years of operational numerical weather prediction reveals much concern with incorporating certain principles of subjective meteorological analysis into the automated procedures. Spatial coherence, temporal continuity, and adherence to dynamic constraints are qualities of subjective analysis that early investigators sought to include in the procedures developed to interpolate randomly-distributed observations to the grid lattice. Consideration of these principles has continued in subsequent years as objective analysis methodology has evolved.

That evolution has been largely driven by two interrelated developments: rapid progress in prediction modeling and advances in observing technology. The first numerical predictions made by the JNWPU were severely limited in domain and in vertical resolution, and the upper air data base consisted of radiosonde stations distributed over Northern Hemisphere continents. Only a few reports from Ocean Station Vessels were available for oceanic regions. In 1980 operational prediction models are global, with much enhanced resolution in both the vertical and horizontal dimensions, and with quite sophisticated modeling of physical processes such as radiation and precipitation. The data base is still founded upon the radiosonde network, but has been augmented by other observing systems such as space-based remote temperature sensors and instruments aboard modern commercial aircraft.

Each of these has had profound impact on the development of methods for providing initial conditions for prediction models. In the succeeding sections of this lecture, these influences are examined within the framework provided by the historical evolution of objective analysis methods at the JNWPU and one of its descendants, the U.S. National Meteorological Center.

II. The Surface-Fitting Method

The first objective analysis method developed by the JNWPU was based on work done by Panofsky (1949) at the suggestion of the meteorology group of the Institute for Advanced Study, Princeton University. Gilchrist and Cressman (1954) describe the method as adapted for JNWPU use. It was tested briefly in July 1955 and found to be too sensitive to erroneous data; after corrective modifications, it was placed in operational use on 10 October 1955 (Staff, JNWPU, 1957).

A local least-squares fit of observations to a second-order polynomial is the basis of the method. Around each grid point where an interpolated (analyzed) value is desired, a local Cartesian coordinate is established, as in Figure 1. Deviations (D_i^n) of the height of an isobaric surface from the standard atmosphere are then expressed as

$$D_i^n = w_o + w_1 x_i + w_2 y_i + w_3 x_i y_i + w_4 x_i^2 + w_5 y_i^2 \tag{1}$$

The coefficients w_j are chosen so as to minimize the mean-square difference E between the D-values given by (1) and those actually observed (D^o) in a least-squares sense:

$$\frac{\partial E}{\partial w_j} = \frac{\partial}{\partial w_j} \sum_{i=1}^{N} (D^o - D^n)_i^2 = 0 \qquad j = 0, 5 \tag{2}$$

This results in a system of six linear equations in the six unknowns, w_j, which can be solved by any of several numerical methods.

The papers by both Panofsky and by Gilchrist and Cressman note that wind information may be allowed to influence the analysis of geopotential height through application of geostrophy. This recognizes that geostrophic wind components u_g and v_g are related to the gradients of D:

$$u_g \sim \frac{\partial D}{\partial y}, \quad v_g \sim \frac{\partial D}{\partial x} \tag{3}$$

Thus, the minimization principle (2) becomes

$$\frac{\partial E}{\partial w_j} = \frac{\partial}{\partial w_j} \left\{ \sum_{i=1}^{N} (D^o - D^n)_i^2 + \sum_{\ell=1}^{M} \left(\frac{\partial D^o}{\partial x} - \frac{\partial D^n}{\partial x}\right)_\ell^2 + \sum_{k=1}^{K} \left(\frac{\partial D^o}{\partial y} - \frac{\partial D}{\partial y}\right)_k^2 \right\} \tag{4}$$

where $\frac{\partial D^o}{\partial x} = \frac{f}{g} v_g^o$ and $\frac{\partial D^o}{\partial y} = -\frac{f}{g} u_g^o$. Equation (4) still results in a sixth-order linear system in the weights w_j. According to eqn. (1), w_o is the desired analysis value of height departure at the grid point (x = y = 0), and in view of (3), w_1 is proportional to the analyzed v-component of the geostrophic wind at the grid point. Likewise, w_2 is proportional to the u-component of the geostrophic wind.

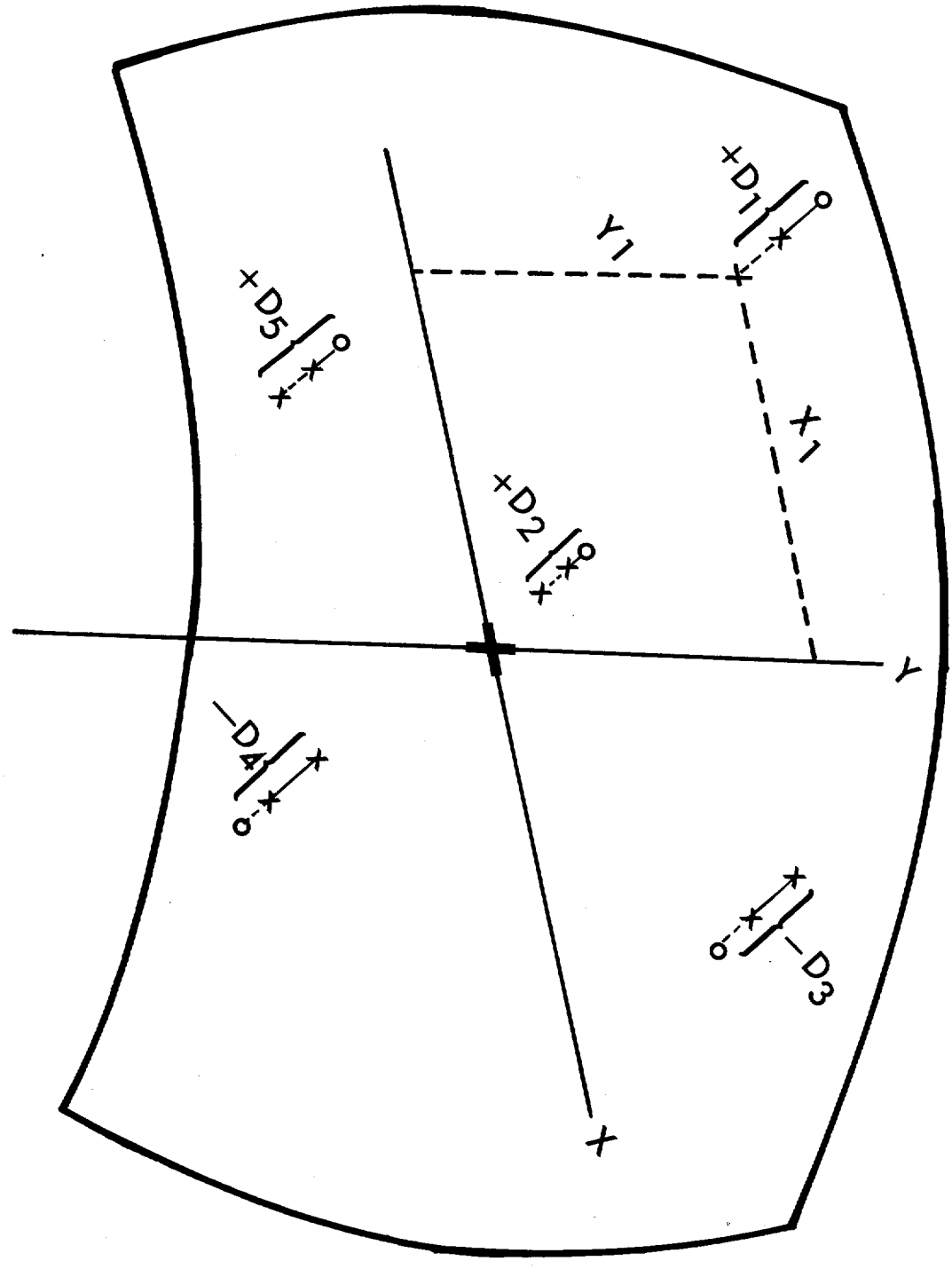

Figure 1. Schematic showing local least-squares fit of observations (D_i) to a quadratic surface (conic section). Algebraic signs represent positions above (+) or below (-) the (x-y) plane. Solid vertical lines denote position above the conic section, and dashed lines represent position below it.

The analysis of both height and wind is therefore obtained by this procedure of fitting the data to a polynomial at each grid point surrounded by sufficient observations (a minimum of six). All of the data used in the fitting procedure receive equal weight; i.e., observations close to the grid point under consideration have the same influence on the analysis as those farther away. Spatial coherence of the analyzed fields depends in part on coherence of the data, in part on data density, and in part on the resolution of the analysis grid. If, for example, the grid points are close enough together and the data are sufficiently dense that part of the data used in the analysis of one grid point is also used in the analysis of an adjacent grid point, then the analyzed fields will not exhibit discontinuities or other unreasonable behavior from point to point.

In practice, the method was applied two-dimensionally on isobaric surfaces with a data base consisting essentially of radiosondes. No effort was made to distinguish data of different qualities, although relative weighting of height data as opposed to wind data was allowed. Vertical structure was depicted as a "stack" of two-dimensional analyses. No explicit attempt was made to ensure vertical consistency, in the sense of adherence to hydrostatic equilibrium or controls on stability, although Gilchrist and Cressman noted in at least one case that a reasonable estimate of vertical stability was calculated from analyzed fields at three levels. As noted above, the method does incorporate a dynamic constraint on the mass and motion fields, in the form of the geostrophic equation. However, there is no application of the principle of temporal continuity: data at a given time were analyzed without knowledge of prior history. Gilchrist and Cressman acknowledge this shortcoming, noting "that the use of a forecast as data or as a background field in data sparse areas....can be compared to the activities of an analyst who consults the previous map when analyzing an area which has poor data coverage." They suggest that such use be made of forecasts in subsequent objective analysis schemes.

To summarize, the first operational JNWPU objective analysis method did attempt to incorporate some of the important principles of subjective analysis--a mass/motion dynamic constraint, for example; and recognized that others, such as temporal continuity, needed inclusion. The method produced analyses with reasonable spatial coherence over data-dense areas, but did poorly elsewhere. Unacceptable behavior in areas where insufficient data transformed the basic problem from interpolation to extrapolation led ultimately to the abandonment of the method in April 1958.

III. The Successive-Corrections Method

Bergthórsson and Döös (1955) proposed an objective analysis procedure which alleviated to a large extent the unreasonable solutions obtained by surface-fitting methods in data-sparse areas. Their method was based on defining corrections to a preliminary or "background" field in such a manner that the correction at a given grid point is a weighted average of the observed minus background differences at all observation points within an "influence radius" about the grid point.

Cressman (1959) adopted this principle of analyzing corrections to a background field, and added the important step of iterating the procedure. In the first iteration, the background field is interpolated by a simple formula to the location of each observation and the difference D between the observed and interpolated values is formed. For a particular grid point where an analysis of geopotential height is desired, a correction is then determined at each observation location within a radius R of the grid point,

$$C_i^h = w_i D_i \tag{5}$$

where $w_i = (R^2 - r_i^2)/(R^2 + r_i^2)$, r_i representing the distance of the i^{th} observation to the grid point being analyzed, as in Fig. 2. The total correction at the grid point is then a linear combination of the individual corrections, each weighted by its distance to the grid point and the total number of corrections being used:

$$D_g = (w_1 D_1 + w_2 D_2 + \ldots + w_n D_n)/n \tag{6}$$

where D_g is the analyzed correction at the grid point. Two points are worth noting: first, because the weight accorded any individual correction is normalized by the total number of corrections, the magnitude of D_g cannot exceed the magnitude of any individual correction. Second, for the special case of coincident observation location and grid point, r = 0, and the weight given the correction is 1/n. If this is the only report considered, its weight is unity; that is, the analysis exactly reflects the data.

The result of adding the corrections given by eqn. (6) to the background field forms a new background field, which reflects the data to some degree. New differences between observed and interpolated reports (presumably smaller) may then be determined. The correction process embodied in eqns. (5) and (6) is then repeated, completing a second iteration through the data. In practice, four iterations have been found to produce satisfactory results. The influence radius is decreased in each successive scan to better reflect small-scale features.

As in the surface-fitting method, wind information is allowed to influence the analysis of geopotential height by assuming a geostrophic relationship[1] between the mass and motion fields. If a wind is available, a correction to the height field is defined by

$$C_j^v = w_j[D_j^o + K(u_j^o \Delta x - v_j^o \Delta y) - D_g^*] \tag{7}$$

---

[1]An approximate gradient wind relationship was introduced in the middle 1960's.

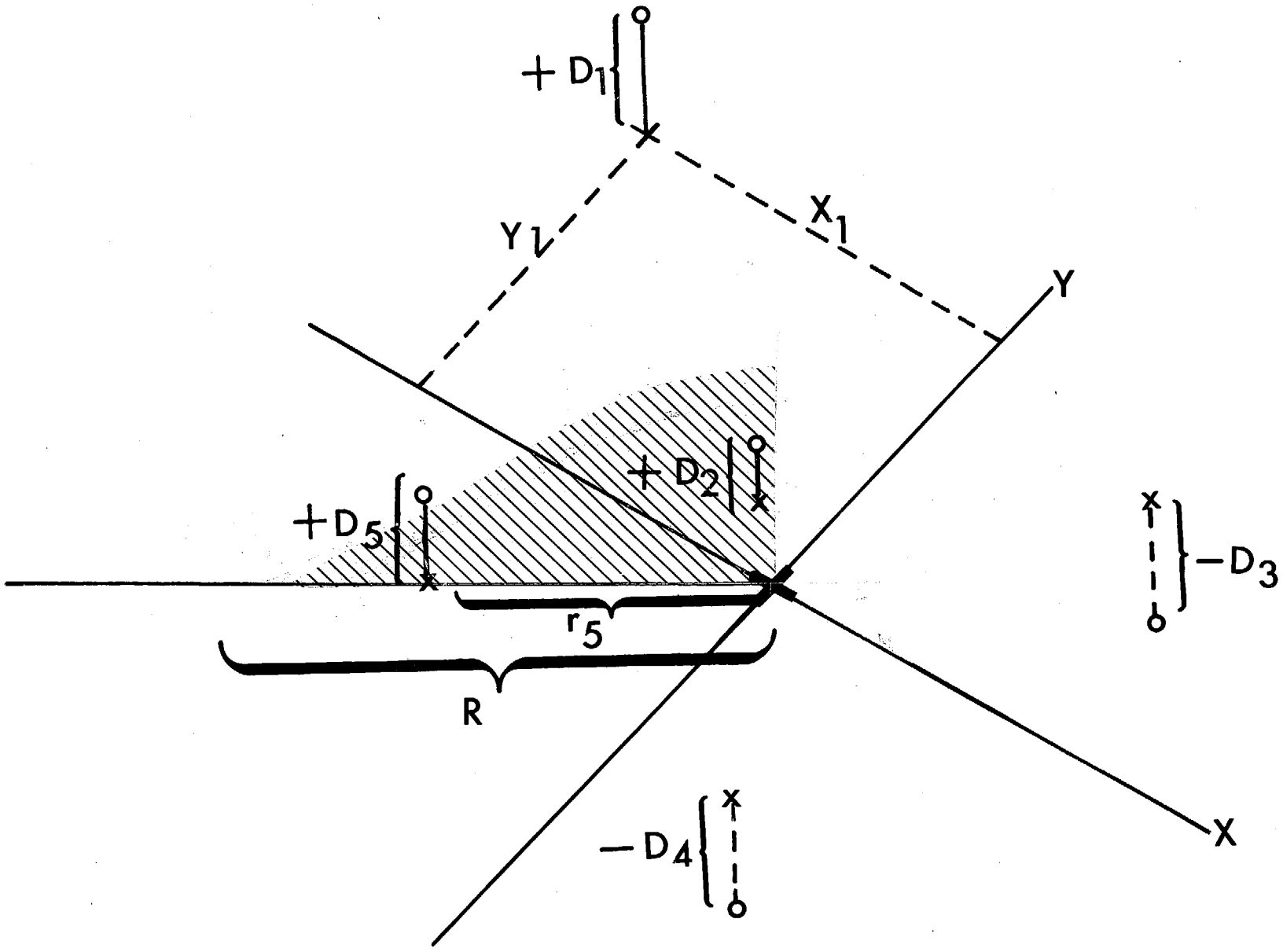

Figure 2. Distribution of observations as in Fig. 1, but with the successive-corrections weighting function superimposed.

where $K = \frac{f}{mg}\frac{U_g}{U}$, with U_g/U the average ratio of geostrophic to actual wind speed, u_j^o and v_j^o the observed wind components at a distance $r_j = [(\Delta x)_j^2 + (\Delta y)_j^2]^{\frac{1}{2}}$ from the grid point. D_j^o is the observed height at observation point j, and D_g^* is the background value at the grid point. The total correction to the geopotential height at the grid point is then

$$C = \frac{1}{n_h + n_v}\left\{\sum_{i=1}^{n_h} C_i^h + \sum_{j=1}^{n_v} C_j^v\right\} \qquad (8)$$

where n_h is the number of height observations and n_v is the number of wind observations. This procedure essentially extrapolates the geopotential gradient (implied by assuming the observed winds are geostrophic) to obtain an estimated geopotential at the grid point. When observed winds are used in this manner, it is no longer possible to guarantee that the analyzed correction at a grid point cannot exceed any individual height correction.

As used operationally, wind observations are allowed to influence the height analysis in approximately the manner outlined above. The height analysis is performed two-dimensionally, using a 12h forecast height field as the background field. Analyses at 1000-mb and 300-mb are performed first. Linear combinations of these analyses are then formed to serve as background fields for intermediate levels. Statistical regression is used to infer background fields above 300 mb from the 300 mb analysis. Thus, although the interpolation method itself is two-dimensional, the sequence of application exerts a strong tendency for consistent vertical structure. The use of a forecast as a background field represents a consideration of temporal continuity in the analysis, and contributes to greater spatial coherence. The latter is also enhanced by the more stable character of the interpolation method itself.

An analysis of scalar wind components is also performed. Here the background fields are obtained from the height analysis by means of the balance equation. Corrections of the form of eqn. (5) are then made based on observed wind components. Thus, the analysis of heights reflects available winds to some extent, and the analysis of winds reflects height information. Mass and motion analyses are therefore related, at least to some degree, via a dynamic constraint.

The successive correction method was placed in operational use in April 1958. By the early 1970's the method was universally employed in all NMC objective analyses: a limited-area version provided initial conditions for the limited-area, fine-mesh (LFM) model (Gerrity, 1977), and hemispheric versions supported the hemispheric prediction model (Shuman and Hovermale, 1968) and provided a late, or "final" analysis for each synoptic time. In September 1974, it was supplanted in the latter two uses by a global analysis method to be briefly described in the next section. It remains in use for support of the LFM.

IV. Spectral Objective Analysis Method

By the late 1960's a number of developments had combined to allow the expansion of operational numerical weather prediction to the entire globe. Finite-difference representations appropriate to spherical geometry had been developed and applied to large, sophisticated prediction models in general circulation research, showing that models could be integrated more-or-less properly in time. Of equal importance, remote sensing technology combined with space-based observing platforms suggested for the first time that eventually the atmosphere could be observed on a global basis.

Development therefore began at NMC in the late 1960's on a global prediction model suitable for operational use. In parallel, development of a suitable analysis method also began in October 1968 with the arrival at NMC of then Captain T. W. Flattery, on assignment as U.S. Air Force Liaison Officer.

Flattery (1971) devised a surface-fitting method whose antecedents include the method of Gilchrist and Cressman, and the extensions to three space dimensions by Corby (1963) and Dixon et al. (1972). Unlike its predecessors, Flattery's spectral objective analysis is based on a least-squares fit of available data to a family of surfaces defined globally rather than locally. It is unique in that the surfaces are defined in part by Hough functions, eigenfunctions of Laplace's tidal equation. Because these functions represent some of the natural oscillatory modes of the atmosphere, their use lends a physical basis to surface-fitting objective analysis, as opposed to the purely mathematical basis of the earlier methods.

The analysis method expresses geopotential height (z), zonal wind (u), and meridional wind (v) as functions of latitude (ϕ), longitude (λ), and pressure (p):

$$z(\lambda,\phi,p) = \sum_{\ell=0}^{24}\sum_{m=1}^{24}\sum_{n=1}^{7}\{[a_{\ell,m,n}\cos(\ell\lambda) + b_{\ell,m,n}\sin(\ell\lambda)]\, H_{\ell,m}(\phi)\, E_n(p)\}$$

$$u(\lambda,\phi,p) = \sum_{\ell=0}^{24}\sum_{m=1}^{24}\sum_{n=1}^{7}\{[a_{\ell,m,n}\cos(\ell\lambda) + b_{\ell,m,n}\sin(\ell\lambda)]\, U_{\ell,m}(\phi)E_n(p)\} \quad (9)$$

$$v(\lambda,\phi,p) = \sum_{\ell=0}^{24}\sum_{m=1}^{24}\sum_{n=1}^{7}\{[a_{\ell,m,n}\sin(\ell\lambda) - b_{\ell,m,n}\cos(\ell\lambda)]\, V_{\ell,m}(\phi)E_n(p)\}$$

The $E_n(p)$ are empirical orthogonal functions used to represent vertical structure; $H_{\ell,m}(\phi)$ are Hough functions; $U_{\ell,m}(\phi)$ and $V_{\ell,m}(\phi)$ express latitudinal variations in velocity components. The wind functions are derived from the Hough functions such that U, V, and H are related in a quasi-geostrophic manner.

The coefficients a and b are determined by subdividing the atmosphere from the surface to 50mb into N volume elements ( ) and minimizing the least-squares summation

$$S = \sum_{i=1}^{N} \{\xi[\overline{z}_{ob})_i - (z_{an})_i]^2 + [(\overline{u}_{ob})_i - (u_{an})_i]^2 + [(\overline{v}_{ob})_i - (v_{an})_i]^2\}\Delta r_i \qquad (10)$$

where ξ is a weighting factor which allows relative weighting of height observations against observations of wind. The quantities ($\overline{z}_{ob}$, $\overline{u}_{ob}$, $\overline{v}_{ob}$) represent the weighted average of observed values of height and wind in each volume element; (z_{an}, u_{an}, v_{an}) represent analyzed values. The analysis consists of finding coefficients $a_{\ell,m,n}$, $b_{\ell,m,n}$ such that the summation S is a minimum.

Horizontal spatial coherence is ensured by the global definition of the basis functions. The empirical orthogonal functions used to represent vertical structure embody the hydrostatic constraint through the observational data they are derived from. Consequently, analyses obtained by least-squares fit to such functions also tend to exhibit hydrostatic equilibrium.

Temporal continuity is partially achieved through use of a forecast as a background. Like the successive corrections method, the procedure is iterative: the coefficients representing the background, or "guess" field are modified by successive scans through the data. Resolution is increased during each scan. During the first iteration, the guess coefficients of the terms of the series representation involving the lowest four vertical modes, the lowest seven Hough functions, and the lowest ten longitudinal modes are modified to reflect the available data. These modified coefficients serve as "guesses" for the second iteration, during which additional modes are included. Nine scans through the observations are performed, the last four at full horizontal resolution. Experience has shown that nine scans through the data are usually sufficient for convergence to a stable solution. Full resolution is 7 vertical modes, 24 latitudinal modes, and 24 longitudinal modes.

In the analysis, heights and winds are analyzed simultaneously. The degree to which the wind law implicit in the relationship between U, V, and H is enforced may be controlled experimentally through the choice of ξ in Eqn. 10. In practice, the summation is minimized twice, once with a value of ξ appropriate for the height analysis and once with a value

appropriate for the wind analysis. This procedure produces two sets of analysis coefficients: one for heights, the other for winds. Numerical values of ξ are chosen so that heights receive more weight than winds in the height analysis, winds more weight than heights in the wind analysis.

By 1972, the spectral objective analysis method was undergoing quasi-operational testing. Its implementation into actual NMC operations was delayed because of a major change in computers until September 1974. It then replaced the successive-corrections method in providing initial conditions for the hemispheric prediction model and a "final" analysis. The former application is still operational; for the latter, the "final" analysis was replaced in September 1978 by a data assimilation system based on a statistical analysis method called "optimum interpolation." Before describing that method, it is necessary to consider the concepts which are somewhat loosely bound together under the label "data assimilation."

V. Data Assimilation

All of the objective analysis systems previously described were designed for an upper air data base consisting mostly of radiosonde reports of temperature and wind profiles. With the launching of artificial earth satellites in the late 1950's, observations of the global atmosphere from another source--remote sensing--became a practical possibility. Inversion of the radiative transfer equation to obtain temperature profiles had been studied as a theoretical problem beginning early in the 20th century (Wark and Fleming, 1966). Less than 2 years after the launch of the first artificial satellite, Kaplan (1959) proposed to transform the problem from the theoretical to the practical. By early 1969, remotely-sensed temperature soundings were being processed quasi-operationally and made available to NMC for examination. Admission of these data to the operational data base occurred on 28 May 1969.

During the decade preceding operational implementation, certain questions raised by the nature of remote sounding data came under consideration. An obvious difference from radiosondes is that the remote soundings are incomplete: they provide information only on the mass field. A second difference arises from the nature of the observing platform, a polar-orbiting satellite. Soundings are available along and some distance either side of the path of the satellite as it moves in orbit. Thus the sounding data are not located at fixed points in space or at fixed (synoptic) times, but provide in principle a time history of the thermal field as the satellite moves.

The question of whether the time history of the temperature afforded by remote soundings could compensate for their incompleteness was addressed by Charney, Halem, and Jastrow (1969). Their short paper is generally regarded as the start of an era of vigorous research in what later came to be called "data assimilation." The concepts originating in this research have subsequently had considerable influence on the evolution of numerical weather analysis and prediction.

Charney et al. conducted a series of experiments in which the true atmosphere was simulated by an extended integration of a general circulation model. A time sequence of "observed" temperatures was extracted from this integration to simulate data. Then, to simulate our imperfect approximations to truth, the prediction model was re-integrated beginning several days prior to the start of the sequence of pseudo-data from an initial state perturbed by random errors of temperature. After 10 days of integration, during which period the perturbed integration deviated increasingly from the "truth," the correct temperatures--the pseudo-data--were introduced at intervals of a few hours for an additional 60 days. Several such idealized experiments were performed in which different aspects were varied, including adding random errors to the "observed" temperatures, and decreasing the insertion interval from 12 hours to 1 hour.

Figures 3 and 4, taken from Charney et al., illustrate the main results. In middle latitudes, the curve of Fig. 3 suggests that the motion field can indeed be inferred solely from a time history of the mass field, but that the process of induction is quite slow. Approximately 20 days of integration, inserting "observed" temperatures each 12 hours, was required to reduce the error in the induced wind to an asymptotic value. The asymptotic level of wind error varied in proportion to the assumed error level of the temperature pseudo-data. Figure 4 presents the same curve for the tropics, and shows that the process of inducing the motion field from the mass field proceeds even more slowly and is more sensitive to errors in the mass field.

Thus, the simulation experiments of Charney et al. demonstrated that the motion field can be inferred from a time history of the mass field in principle. In practice, of course, one cannot wait 20 days for the motion field to respond to an observed change in the mass field. Furthermore, the assumptions concerning the error structure of the mass observations have proven quite unrealistic. Subsequent research has also shown that the "identical-twin" approach in simulation experiments--using the same prediction model to represent both the "true" atmosphere and the imperfect model atmosphere--leads to unduly optimistic conclusions.

Even though some of the conclusions have proven to be unrealistic, the paper of Charney et al. has been very influential. The emphasis it placed on the role of sophisticated prediction models in deducing the current state of the atmosphere from a data base which includes incomplete observations can clearly be identified in the subsequent data assimilation literature (reviewed by Kasahara (1972), Bengtsson (1975), and McPherson (1975)). Most papers envision a prediction model marching forward in time, periodically (or perhaps continuously) being corrected (or "updated") by observations which are in general asynoptic and incomplete. Variables which are not observed are to be inferred from those which are, subject to the constraints embodied in the prediction model equations. This differs from the methods of objective analysis discussed in previous

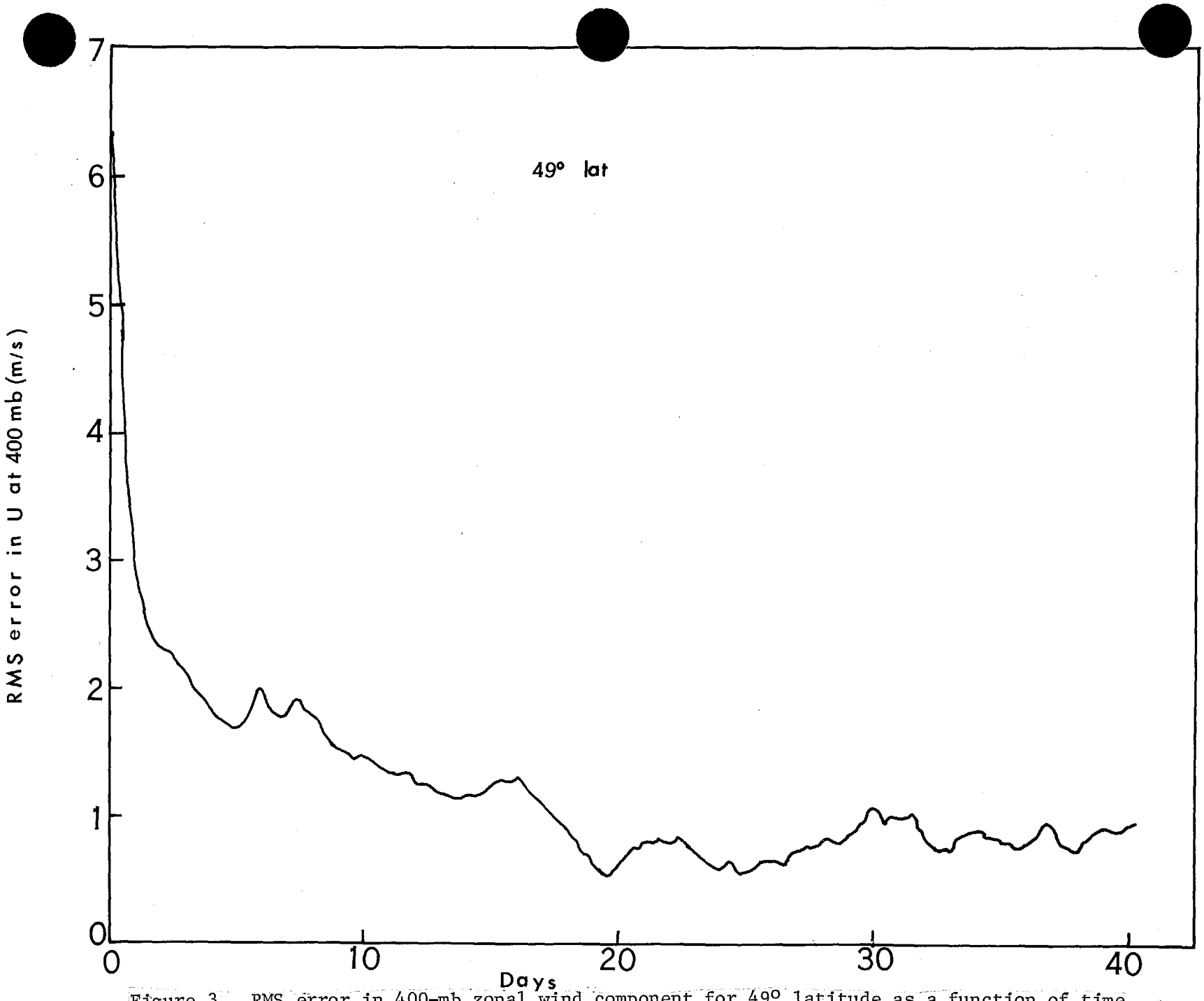

Figure 3. RMS error in 400-mb zonal wind component for 49° latitude as a function of time. Wind induced by inserting "perfect" simulated temperatures plus random error of 1°C. After Charney, Halem, and Jastrow (1969).

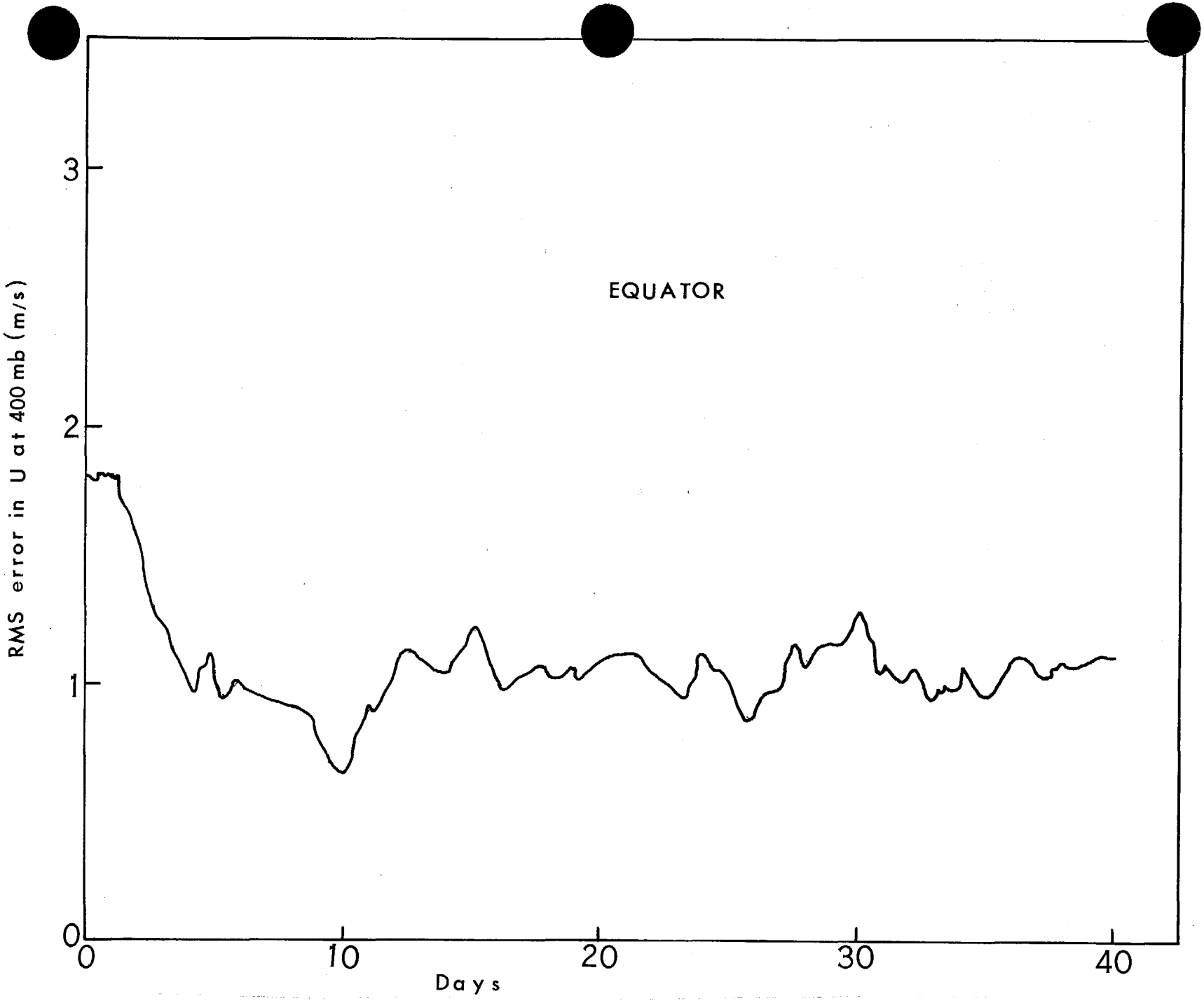

Figure 4. Same as Fig. 3, but for the equator. After Charney, Halem, and Jastrow (1969).

sections mostly in the greater generality of the dynamic constraints. The emphasis which this places on the role of the prediction model as the primary vehicle for representing the atmosphere is one of the main consequences of the decade of data assimilation research.

Another is that observations of one variable alone are not sufficient, because of observational and modeling errors. It became clear that the global observing system would not continue as a homogeneous network of conventional land upper air networks, as in previous years. Rather, it would be a mixture of in situ and remotely-sensed measurements, some measuring mass, others motion, and all with differing error characteristics. This nonhomogeneous data base could not be treated easily by existing objective analysis methods, and so was a major factor in the design of a new system, described in the next section.

VI. Multivariate Statistical Analysis Method

In the early 1960's, Gandin (1963), and independently Eddy (1967), developed an analysis procedure based on statistical interpolation. According to an adaptation of this method by Bergman (1979), the analyzed value at any point is related to observations in its vicinity by

$$F_g^a = F_g + w_1(F^o - \hat{F})_1 + w_2(F^o - \hat{F})_2 + \cdots w_n(F^o - \hat{F})_n, \qquad (11)$$

where F_g^a = analyzed value at the point g;

$\hat{F}_g$ = forecast value at the point g;

$(F^o - \hat{F})$ = difference between the observation (F^o) and the forecast interpolated to the observation location.

The analysis consists of determining the weights w_i such that the expected mean-square error of interpolation, E^2, is minimized:

$$\frac{\partial E^2}{\partial w_j} = \frac{\partial}{\partial w_j} \overline{\left(F_g^a - \hat{F}_g - \sum_{i=1}^{n} w_i f_i\right)^2} = 0 \qquad (12)$$

where $f_i = (F^o - F)_i$ represents the error in the forecast at point i. Performing the indicated differentiation in eqn. (12) with respect to each of the w_i, and equating the results to zero yields a system of n linear equations of the form

$$\begin{bmatrix} (\sigma_f^2 + \overline{\varepsilon_1^2}) & (\overline{f_1 f_2} + \overline{\varepsilon_1 \varepsilon_2}) & (\overline{f_1 f_3} + \overline{\varepsilon_1 \varepsilon_3}) \cdots & (\overline{f_1 f_n} + \overline{\varepsilon_1 \varepsilon_n}) \\ (\overline{f_2 f_1} + \overline{\varepsilon_2 \varepsilon_1}) & (\sigma_f^2 + \overline{\varepsilon_2^2}) & (\overline{f_2 f_3} + \overline{\varepsilon_2 \varepsilon_3}) \cdots & (\overline{f_2 f_n} + \overline{\varepsilon_2 \varepsilon_n}) \\ \vdots & \vdots & \vdots & \vdots \\ (\overline{f_n f_1} + \overline{\varepsilon_n \varepsilon_1}) & + \cdots & \cdots & (\sigma_f^2 + \overline{\varepsilon_n^2}) \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} \overline{f_g f_1} \\ \overline{f_g f_2} \\ \vdots \\ f_g f_n \end{bmatrix} \qquad (13)$$

where $\overline{f_i f_j}$ represents the covariance of the forecast error at point i with that at point j, and $\overline{f_i f_g}$ represents the covariance of the forecast error at point i with that at the grid point. If the several covariances are known, then the system (13) can be solved for the weights w_i.

It is important to note that the observations are always in error to some degree, so that the observed forecast error is composed of the true forecast error plus an apparent error due to inaccurate observations:

$$f_i = f_i(\text{true}) + \varepsilon_i \qquad (14)$$

Thus the forecast error covariance $\overline{f_i f_j}$ may be written as

$$\overline{f_i f_j} = \overline{f_i(\text{true}) f_j(\text{true})} + \overline{\varepsilon_i \varepsilon_j} + \overline{f_i(\text{true})\varepsilon_j} + \overline{f_j(\text{true})\varepsilon_i}. \qquad (15)$$

The last two terms on the right side of eqn. (15) represent the covariance of the observational error at one point with the true forecast error at another point. These are usually assumed to vanish. The second term on the right represents the covariance of the observational error at one point with that at another point. For random errors, this term vanishes except when i=j; then it represents the observational error variance.

It is this term which permits the statistical analysis method to discriminate among data with differing error characteristics. Basically, a large value of $\overline{\varepsilon_i \varepsilon_i}$ for a particular class of data will result in smaller values of the weights w_i in eqn. (11) for that class, compared to other data sets. Thus, if the error characteristics of the different members of the nonhomogeneous data base are known, the statistical analysis method can effect a statistically optimum blend.

In practice, the covariance function $\overline{f_i f_j}$ has been modeled by an analytic function to fit observed forecast errors in geopotential height. It will be shown in the next lecture of this series that, under the assumption that forecast errors in winds and temperatures are related to forecast errors in heights through the geostrophic and hydrostatic relationships, one can derive all other auto-covariances and cross-covariances necessary to allow motion data to enter the mass analysis and vice versa.

Leaning on the experience derived from data-assimilation research, the statistical analysis method is applied in a way which emphasizes the importance of the prediction model. Several key design features, new to analysis practice at NMC, resulted:

- the analysis consists of correcting the model locally with available data, in the terrain-following σ-coordinate of the prediction model rather than on isobaric surfaces;

- the variables updated are the history variables of the model--winds, layer-mean temperature, surface and tropopause pressures, and specific humidity--in part to tailor the updating method to the prediction model as closely as possible, and in part to acknowledge that remote sounding methodology produces temperature estimates from deep layers;

- observations and first guess are weighted in the analysis according to estimates of relative accuracy.

The statistical analysis method thus incorporates the main principles of subjective analysis its predecessors had also incorporated; spatial and temporal coherence through use of a prediction as a background field, and adherence to simple dynamic constraints. In addition, it builds upon the main results of data assimilation research.

A global data assimilation based on this method has been operational since September 1978 (McPherson et al., 1979). Its performance has been such as to warrant consideration for all analysis and assimilation functions at NMC. Consequently, two full lectures of this series are devoted to discussing the theoretical foundation of the method and its practical application. At present, however, NMC is in a time of transition, and the extant analysis/assimilation system is a mixture of methods.

VII. Present Operational Assimilation and Analysis Systems

The present NMC system for operational numerical weather analysis and prediction features a global data assimilation system and three analysis/forecast systems. Figure 5 schematically shows the present structure. At either primary synoptic time, three global analyses are performed. All update the same prediction, the 6h global forecast from the assimilation system. The first is initiated only 80 minutes after 00 GMT or 12 GMT, and provides initial conditions for the "early" integration of the barotropic model. Presently, this update uses the spectral objective analysis method.

At roughly 3.5h after 00 GMT or 12 GMT, the second global analysis is begun. No knowledge of the "barotropic" analysis is forwarded to the second, or "large-scale" analysis. All data that were available to the first analysis, plus observations arriving at NMC between 1:20 and 3:30, are available to the second one. As in the barotropic cycle, this analysis is performed by the spectral objective analysis method.

The third global analysis occurs in the assimilation cycle, initiated approximately 9.5h after 00 GMT and 12 GMT. At present this is done in two segments. First is the final update of the main synoptic time, followed by a 6-hour prediction. The second segment consists of updating that prediction using observations centered around 06 GMT or 18 GMT (data cutoff of roughly 4h) and then a 6h prediction valid at the next primary observation time.

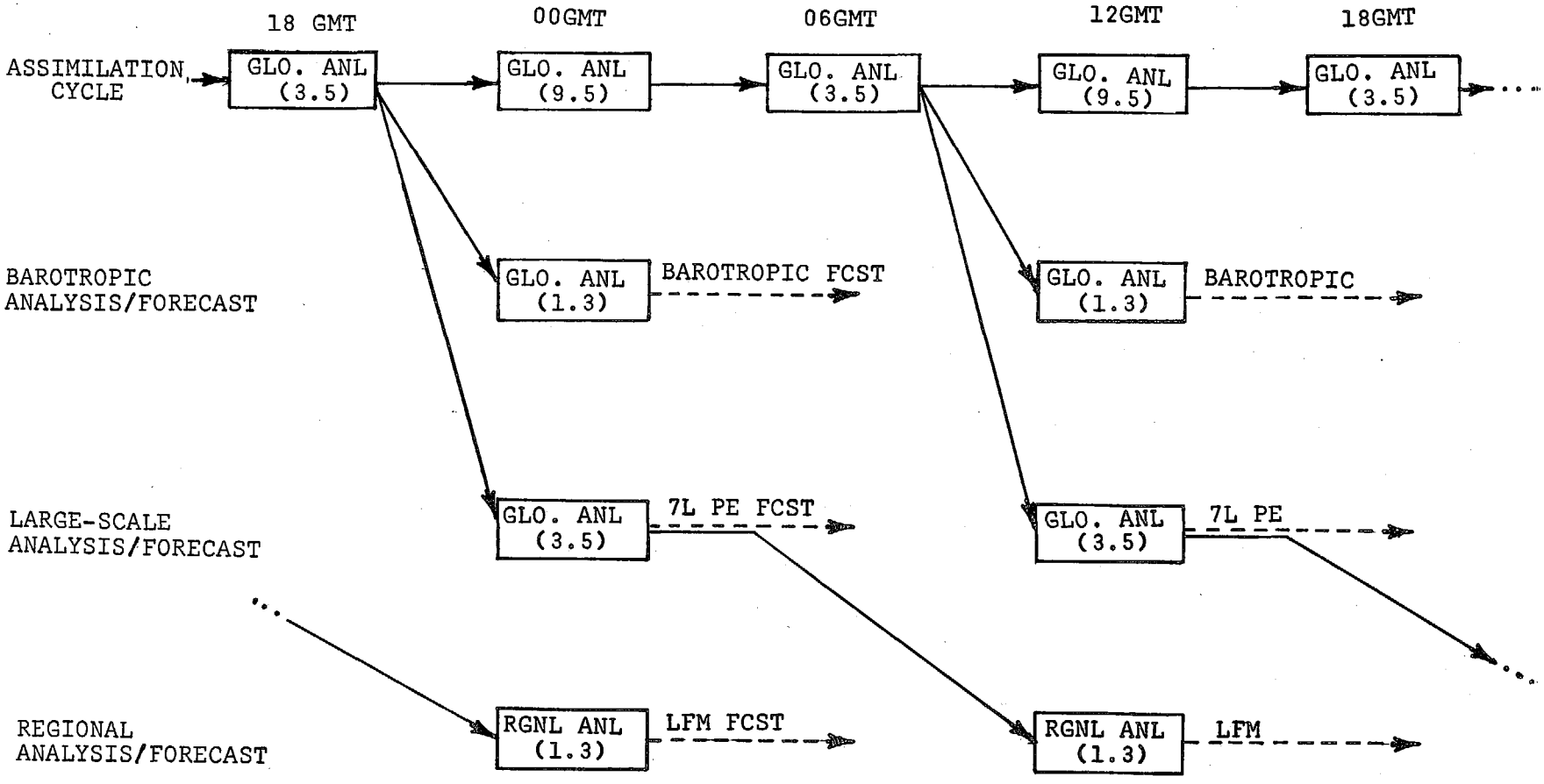

Figure 5. Schematic diagram of the NMC operational assimilation and analysis/forecast system as of 1 January 1980. Boxes represent analyses (updates), with parenthetical numbers indicating cutoff time. Solid lines denote the origin of the background fields for each update, and the dashed lines represent operational forecasts.

Several features of this system are worth emphasizing. First, the assimilation system is separate and distinct from the analysis/forecast system used to produce numerical forecasts. The latter takes its first guess from the former, but there is no other connection. Both prediction models and updating or analysis methods are different: in the assimilation cycle, the nine-layer latitude-longitude model is updated by optimum interpolation, but in the analysis-forecast system the analysis is the Hough spectral method and the model is a hemispheric, seven-layer model using a polar-stereographic projection. The characteristics of these elements differ considerably.

Second, the large-scale update (3.5h) has no knowledge of the barotropic (1.3h) update; it merely repeats the work of its predecessor, and takes no advantage of the fact that some observations have been considered twice. Similarly, the assimilation update knows nothing of the first two, repeating their work, and perhaps their mistakes. In the present system, then, three passes are made through at least some of the data, and the large majority are scanned twice but little or no profit is realized from the possibilities of iteration.

The regional component of the present system is also separate and distinct. Its analysis is the successive approximation method, although it has undergone many modifications since becoming operational more than two decades ago. The regional prediction model is a high resolution version of the Shuman-Hovermale seven-layer model used in the large-scale cycle but modified for limited-area integration. It also has subtle differences in other aspects, such as precipitation modeling.

At present, the regional analysis takes as its first guess the 12th hour of the large-scale prediction from the previous cycle. It has knowledge of the assimilation cycle only indirectly, through the spectral analysis and subsequent large-scale prediction. The regional analysis therefore is not influenced by late-arriving data or data at 06 GMT and 18 GMT, which enter the assimilation system.

Table 1 summarizes the characteristics of the three main analysis methods as they are currently applied.

VIII. Summary

A highly condensed account has been presented of the evolution of objective analysis and data assimilation methods at NMC and its antecedent organizations. The discussion has used as a thread of continuity the treatment in each method of three principles of subjective analysis generally thought desirable in objective analysis: spatial coherence, temporal continuity, and incorporation of dynamic constraints.

It was noted that the most recent developments in objective analysis have been motivated by a phenomenal expansion of the data phase and a change in its character. Methods developed earlier for an essentially homogeneous data base are therefore no longer as useful as they once were. The statistical method was developed specifically to treat the existing non-homogeneous data base.

Finally, the assimilation and analysis/forecast systems at NMC will continue to evolve. For example, versions of the statistical analysis method are being prepared for comparison with the successive-correction and spectral methods presently used in the Regional, Large-Scale, and Barotropic runs. Should those statistical analysis methods prove superior in extensive testing, then the diagram of Fig. 5 and the system characteristics in Table 1 will change markedly. Whether or not these particular methods are implemented, change is certain to occur.

Table 1. Characteristics of NMC operational analysis methods as of 1 January 1980.

Characteristic	Analysis Method		
	LFM	Spectral	Statistical
Resolution	191.5 km at 60N	24 modes	2.5^{o} lat-lon surface
Domain	North America	Global	3.75^{o} upper air
First Guess	12h 7L PE	6h 9L Global	6h 9L Global
Variables	z, u, v, RH	z, u, v, RH	p_{σ}, Δz, u, v, q
Interpolation	2-D successive correction (iterative)	3-D spectral (iterative)	3-D statistical
Vertical coordinate	pressure	pressure	sigma
Mass-motion balance	moderately geostrophic	moderately geostrophic	weakly geostrophic
Data selection	all data within influence radius used	all data within 3^{o} lat. x 6^{o} lon. box	10 most highly correlated observations
Data weighting	distance-dependent	two classes	recognizes observational errors

References

Bengtsson, L., 1975: Four-dimensional assimilation of meteorological observations. GARP Publication Series No. 15, Geneva, World Meteorological Organization, 76 pp.

Bergman, K., 1979: Multivariate analysis of temperatures and winds using optimum interpolation. Mon. Wea. Rev., 107, 1423-1444.

Bergthórsson, P., and B. Döös, 1956: Numerical weather map analysis. Tellus, 7, 329-340.

Charney, J., M. Halem, and R. Jastrow, 1969: Use of incomplete historical data to infer the present state of the atmosphere. J. Atmos. Sci., 26, 1160-1163.

Corby, G., 1963: An experiment in three-dimensional objective analysis. Tellus, 15, 432-438.

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

Dixon, R., E. Spackman, I. Jones, and A. Francis, 1972: The global objective analysis of meteorological data using orthogonal polynomial base functions. J. Atmos. Sci., 29, 609-622.

Eddy, A., 1967: The statistical objective analysis of scalar data fields. J. Appl. Meteor., 6, 597-609.

Flattery, T., 1971: Spectral models for global analysis and forecasting. Proceedings, Sixth AWS Technical Exchange Conference, U.S. Naval Academy, Sept. 1970. Air Weather Service Technical Report 242, 42-54.

Gandin, L., 1963: Objective analysis of meteorological fields. Gidrometeorologicheskoe Isdatel'stvo, Leningrad. Translated from Russian, Israel Program for Scientific Translation, Jerusalem, (1965) 242.

Gerrity, J., 1977: The LFM model--1976: a documentation. NOAA Tech. Memo, NWS NMC-60, 68 pp.

Gilchrist, B., and G. Cressman, 1954: An experiment in objective analysis. Tellus, 6, 309-318.

Kaplan, L., 1959: Inference of atmospheric-structure from remote radiation measurements. J. Optical Soc. Amer., 49, 1004-1007.

Kasahara, A., 1972: Simulation experiments for meteorological observing systems for GARP. Bull. Amer. Meteor. Soc., 53, 252-264.

McPherson, R., 1975: Progress, problems, and prospects in meteorological data assimilation. Bull. Amer. Meteor. Soc., 56, 1154-1166.

McPherson, R., K. Bergman, R. Kistler, G. Rasch, and D. Gordon, 1979: The NMC operational global data assimilation system. Mon. Wea. Rev., 107, 1445-1461.

Panofsky, H., 1949: Objective weather map analysis. J. Meteor., 6, 386-392.

Shuman, F., and J. Hovermale, 1968: An operational six-layer primitive equation prediction model. J. Appl. Meteor., 7, 525-547.

Staff, Joint Numerical Weather Prediction Unit, 1957: One year of operational numerical weather prediction. Bull. Amer. Meteor. Soc., 38, 263-268.

Wark, D., and H. Fleming, 1966: Indirect measurements of atmospheric temperature profiles from satellites: I. Introduction. Mon. Wea. Rev., 94, 351-362.