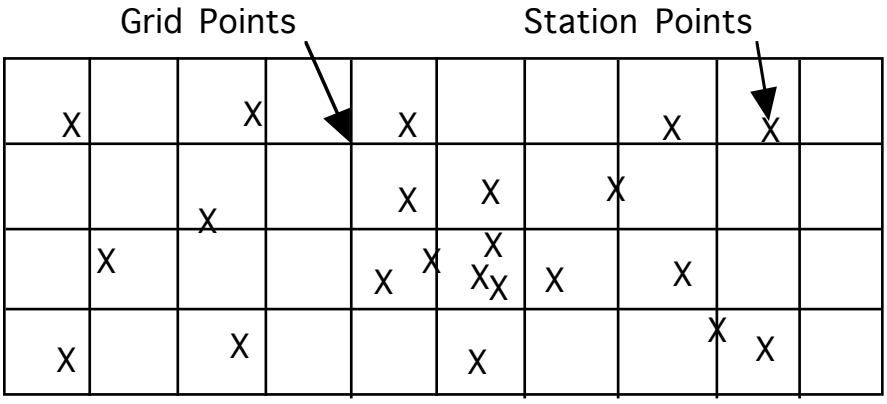
5. Objective Analysis of Observations onto a Regular Grid

In dynamical meteorology, oceanography, and numerical prediction one is often presented with the following problem. Data are available at a number of observation points (usually located near cities or at field stations, along ship cruise tracks, at moorings, or perhaps located by the observation points of an orbiting satellite) that are unevenly distributed over the domain of interest (the globe, for example). In order to compute derivatives of the field variables, as would be required in diagnostic studies or in the initialization of a numerical model, or simply to perform a sensible averaging process, one often requires values of the variables at points on a regular grid. Assigning the best values at the grid points, given data at arbitrarily located stations and perhaps a first guess at regular grid points, is what has traditionally been called objective analysis when done on a computer rather than graphically by hand.



We will use the example of making weather maps from rawinsonde data as the particular example of the mapping problem here. In fact the methods described are applicable to any problem where the data you are given do not fill the domain of interest fully, and/or where the data must be interpolated to a regular grid. The regridding can be in space, in time, or both. You may also find yourself in the position of wanting to plot a continuous function of an observation in two parameter dimensions, and have samples at only a few points. We will proceed through some of the methods in the order that they arose in the history of numerical weather forecasting. In this way we show the weaknesses of some of the most obvious methods such as function fitting, to the correction method, and ultimately to statistically optimized correction methods such as optimum interpolation. Current assimilation schemes in numerical forecast models us a combination of optimum interpolation and use of the governing equations of the model, which we can call Kalman filtering, which is discussed in elementary terms in Chapter 8.

5.1 Polynomial Fitting Methods

Panofsky (1949) suggested a method in which data at station points are used to define the coefficients of a polynomial fit, which is then used to estimate the values of the variables at the grid points. The general form chosen could be,

$$\Phi(x,y) = a_0 + a_1 x + a_2 x^2 + b_2 y^2 + 2c_2 xy + ... \tag{5.1}$$

Polynomial fits are unstable in the sense that the values the polynomials give at points between the stations vary greatly for small changes in the data at the station points, and especially so when data are missing. The problem gets worse as the order of the polynomial is increased. The method is nearly useless where the data are sparse. The instability of the polynomial fit is such that when one key data point is removed, the polynomial fit in that region may change radically.

Gilchrist and Cressman (1954) suggested a multiple regression to a second-order polynomial, but applied to each individual grid point. A sufficient number of stations in the vicinity of a grid point could be used to determine the parameters of the fit that is appropriate for that grid point. A different set of stations could be used for other fits in the neighborhood of other grid points. If you begin with a form like,

$$\tilde{\Phi}(x,y) = \sum_{k=1}^{6} a_k x^{\alpha_k} y^{\beta_\kappa} \tag{5.2}$$

then wind data can also be included in the estimate of geopotential by using (5.2) and the geostrophic relationship to write,

$$\tilde{U} = -\frac{1}{f_0} \sum_k a_k \beta_k x^{\alpha_k} y^{\beta_\kappa - 1} \tag{5.3}$$

$$\tilde{V} = \frac{1}{f_0} \sum_k a_k \alpha_k y^{\beta_k} x^{\alpha_\kappa - 1} \tag{5.4}$$

A composite error function can be defined, that can be minimized to define the coefficients, a_k.

$$E = \sum_{i=1}^{N} \left(\tilde{\Phi}_i - \Phi_i \right)^2 + \lambda^2 \left\{ \sum_{i=1}^{N} \left(\tilde{U}_i - U_i \right)^2 + \sum_{i=1}^{N} \left(\tilde{V}_i - V_i \right)^2 \right\} \tag{5.5}$$

Here the subscript *i* refers to the stations where we have data. Of course we must have a minimum of 6 observations to define 6 coefficients. λ^2 is a weight factor to determine how closely to fit to the wind data relative to the height data. This method is relatively expensive and does poorly in regions of sparse data, where we need an objective analysis scheme the most. In this respect it is similar to other polynomial schemes.

The problem with polynomial fits in regions of sparse data is illustrated in the accompanying figure. Shown at the top is the fit of a cubic to five data points. At the bottom we remove one of the data points and do the fit again. Note how wildly the two curves depart from each other in the vicinity of the missing point. Such problems can be avoided by not using polynomial fits, and by utilizing a first guess that retains the information gained from prior observations. Then one missing datum in the middle of the Pacific will not have such an unfortunate effect on the analysis.

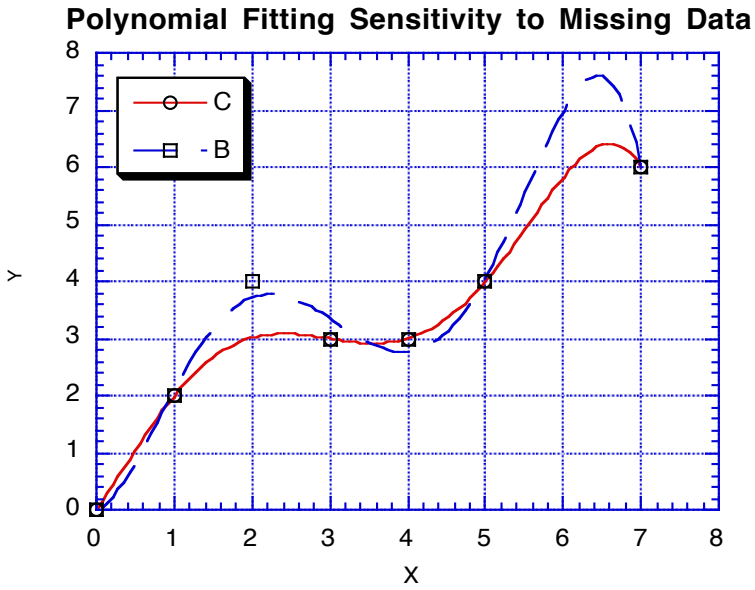


Figure 5.1 . Example of fifth order polynomial fit to data. When data point at x=2 is added (curve B), the fitted curve deviates not only locally at x=2, but also at 5<x<7, where no new data were added.

A polynomial fit that actually got adopted by the US National Meteorological Center for its routine operational products was proposed by Flattery (1971). In this scheme Hough functions were used as the interpolating polynomials. These functions are an orthogonal set that are the solutions of the linearized equations for a resting atmosphere (the tidal equations). The idea was that if you expressed the data in terms of actual solutions of the dynamical equations, then your fit between the data points would have some dynamical consistency. The Hough functions are global functions and so all of the observations were used simultaneously to define the global Hough function

coefficients and produce a global map. Only the Hough functions describing slowly varying rotational modes were used. The gravity wave modes were zeroed out to produce a well-initialized field. This method replaced Cressman's correction method for global analyses in about 1972 and was replaced by Optimum Interpolation in 1978. This method has some dynamical and mathematical appeal, but is in truth just a glorified polynomial fit and has all of the problems of polynomial fits. First of all, the atmosphere is highly nonlinear and strongly forced by heating, especially in the tropics. The Hough modes chosen were primarily the free, non-divergent Rossby modes, which constitute a large, but not dominant, fraction of the variance. Therefore this aspect of the Flattery method did not buy much. In the tropics, where highly divergent motions forced by heating are important, the analyses constructed with the Flattery method are very much in error, especially in their estimates of divergence, which they set to essentially zero. In addition the Hough function fits are wildly unstable in regions of sparse data, like any polynomial fit. The NMC tropical analyses produced before 1978 are almost totally useless because they were made with the Flattery analysis system. Normal mode fits are still used in numerical initialization schemes to remove fast gravity waves, but this does not really affect the slowly changing meteorological flow. Modern reanalysis data products are based on data assimilation methods that take into accout both the data and the model forecast and the uncertainty in both.

5.2 The Correction Method

The correction method (Berthorsson and Doos, 1955; Cressman, 1959), begins with a first guess and modifies it appropriately to take into account new data. Starting with a reasonable first guess and only modifying it when and where data are available can avoid the wildness and instability of the polynomial methods. Also, if the new data departs too wildly from the first guess, one suspects that the data are faulty and can devise objective methods to reject it. The method proceeds by scanning and correcting the field several times and applying some smoothing between correction steps.

Suppose we consider first only the height field. Define the following variables and symbols:

z_f = the first guess at a grid point

z_{oi} = an observation at station i

z_{fi} = the first guess interpolated to the position of station i

The error in the first guess at the station i is then

$$E_h = z_{fi} - z_{oi}$$

The correction required at the grid point is then defined to be

$$C_h = -W E_h$$

Where W is a weighting factor that depends on the distance from the station to the grid point where we are correcting our analysis.

$$W = \frac{D^2 - d^2}{D^2 + d^2}$$

Where d is the distance from the grid point to the station and D is the distance at which $W \to 0$. D is the influence radius.

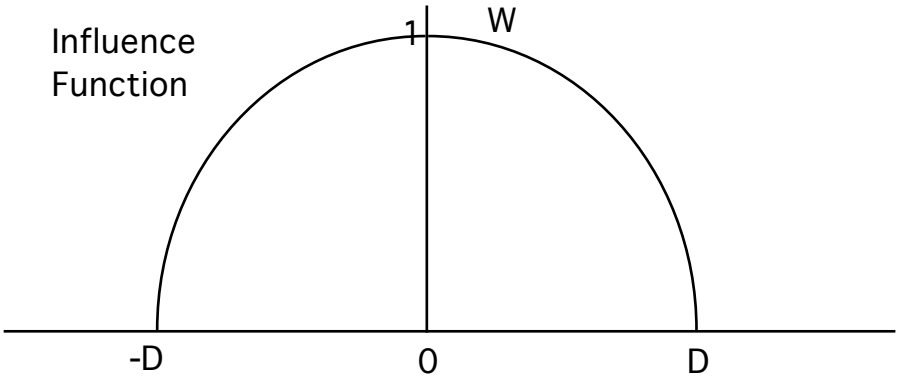


Fig. 5.2 A circular influence function as used in the Cressman scheme.

If both a wind and a height are reported at the station then we can use the geostrophic relationship to extrapolate the observed height to the grid point.

$$C_V = W\left[z_{0i} + \left\{\frac{\partial z}{\partial x}\right\}_i \Delta x + \left\{\frac{\partial z}{\partial y}\right\}_i \Delta y - z_f \right] \tag{5.6}$$

The first three terms inside the square brackets in (5.6) are the observed height at the station i extrapolated to the grid point. If we subtract the first guess at the grid point, z_f, then we have minus the error at the grid point, which if added to the first guess should give an improved estimate of the true height at the grid point. The horizontal derivatives of the height field at the station point can be estimated from the observed winds using the geostrophic relation.

$$\frac{f\, v_g}{g} = m\frac{\partial z}{\partial x}; \qquad \frac{f\, u_g}{g} = -m\frac{\partial z}{\partial y} \tag{5.7}$$

The resulting form of the correction using the winds is then,

$$C_v = W\left[z_{0i} + \frac{kf}{mg}\left(v_i \Delta x - u_i \Delta y\right) - z_f \right] \tag{5.8}$$

Where m is a map factor and $k = u_g/u = 1.08$, on average.

In Cressman's scheme both C_h and C_v are used. Moreover, all station data within a distance D of the gridpoint are used. The actual correction used is a weighted mean of all the C_h and C_v within a distance D of the gridpoint.

$$C = \frac{A\sum_{i=1}^{N_h} C_h + \sum_{i=1}^{N_v} C_v}{A N_h + N_v} \tag{5.9}$$

Here N_h and N_v are the total number of height-based and wind-based corrections, respectively. A is the weighting of height corrections relative to those based on horizontal gradients. Typically A is about 0.25.

One can scan the domain successively, using the previous estimate as the first guess for the new analysis. One would start with a rather large D to make small corrections in the large gaps between stations and then reduce the value of D so that more spatial resolution could be retained in those regions with densely spaced stations. Each scan and correction of the field is best followed by a smoothing operation to remove any kinks that have been introduced by the correction scheme. This method is slow, but relatively successful. It was used by NMC prior to the introduction of the Flattery analysis scheme in 1972, and was called Cressman's scheme.

5.3 Optimum Interpolation

'The interpolation which is linear relative to the initial data and whose root-mean-square error is minimum is called the "optimum" interpolation' (Wiener, 1949). The difference between optimum interpolation and linear regression is that the coefficients are not determined anew each time. Gandin (1963) first applied this method most forcefully to the mapping of meteorological data.

Suppose we consider deviations from some "normal" state. This could be climatology or a first guess, depending upon the application.

$$\phi' = \phi - \phi_{\text{norm}}; \quad \phi_{\text{norm}} = \overline{\phi} \text{ or a first guess.} \tag{5.10}$$

Then we try to approximate the value of ϕ at a grid point, ϕ_g, in terms of a linear combination of the values of ϕ at neighboring station points, ϕ_i.

$$\phi_g' = \sum_{i=1}^{N} p_i \phi_i' ; \phi_g' = \text{grid value; } \phi_i' = \text{station values.} \tag{5.11}$$

The coefficients p_i are to be determined by minimizing the mean squared error.

$$E = \overline{\left(\phi_g' - \sum_{i=1}^{N} p_i \phi_i' \right)^2} \tag{5.12}$$

We can write the normalized error as,

$$\varepsilon \equiv \frac{E}{\overline{\phi_g'^2}} = 1 - 2\sum_{i=1}^{N} p_i r_{gi} + \sum_{i=1}^{N}\sum_{j=1}^{N} p_i p_j r_{ij} \tag{5.13}$$

$$\text{where;} \quad r_{gi} = \frac{\overline{\phi_g' \phi_i'}}{\overline{\phi_g'^2}}; \quad r_{ij} = \frac{\overline{\phi_i' \phi_j'}}{\overline{\phi_g'^2}} \tag{5.14}$$

Differentiation with respect to the coefficients leads to the condition of minimization used to determine them.

$$\frac{\partial \varepsilon}{\partial p_i} = -2 r_{gi} + 2\sum_{j=1}^{N} p_j r_{ij} = 0 \qquad i = 1,2,...,N \tag{5.15}$$

Equation (5.15) constitutes a system of N linear equations for the N p's. By substituting the conditions (5.15) into the expression (5.13) for the error, it can be shown that the error obtained after fitting the coefficients is,

$$\varepsilon = 1 - \sum_{i=1}^{N} r_{gi} p_i \tag{5.16}$$

Note: In this simple example, if one of the observation points, k, coincides with a grid point, then $r_{gk} = 1$, and we expect the regression procedure to return $p_k = 1$ and all the other weights zero. In this case the error is zero, $\varepsilon = 0$, since we have assumed the data are perfect. If the station points are uncorrelated with the grid point in question, then $p_i = 0$ and $\varepsilon = 1$, the climatic norm. That is, the error will equal the standard deviation, but no worse.

Observational Error:

In what we have done so far the observations have been assumed to be perfect. Let us now consider what happens if we explicitly take account of the fact that our observations will always contain some error, δ_i.

$$\phi_i' = \phi_{ia}' + \delta_i \tag{5.17}$$

Let's assume, as is usually reasonable, that the error is unbiased (zero mean) and uncorrelated with the true value,

$$\overline{\phi_{ia}{}' \delta_i} = 0$$

and that the errors at the various stations where we have data are uncorrelated,

$$\overline{\delta_i \delta_j} = \begin{cases} \overline{\delta^2} & i = j \\ 0 & i \neq j \end{cases}$$

In this case, rather than (5.13), we obtain,

$$\varepsilon = 1 - 2\sum_i r_{gi} p_i + \sum_i \sum_j r_{ij} p_i p_j + \eta \sum_i p_i^2 \tag{5.18}$$

where r_{ij} = the correlation between the two points

and where η is the ratio of the error variance (noise) to the measurement variance–in other words, the noise-to-signal ratio.

$$\eta = \frac{\overline{\delta^2}}{\overline{\phi_g{}'^2}} \tag{5.19}$$

Minimization of the error leads to the condition,

$$\sum_j r_{ij} p_j + \eta\, p_i = r_{gi} \qquad i = 1,2,3,...,N \tag{5.20}$$

In this case the normalized error is

$$\varepsilon = 1 - \sum_i \sum_j p_i p_j r_{ij} + \eta \sum_i p_i^2 \tag{5.21}$$

What was the effect of including noise in the measurements?

In order to see how optimum interpolation treats the *a priori* information that the observations include some error, or noise, it is instructive to compare the results (5.20) and (5.21) with the results (5.14) and (5.15) obtained assuming perfect data. In the case of perfect data (5.15) gave,

$$r_{ij} p_j = r_{gi}, \qquad \text{or} \qquad p_j = r_{ij}^{-1} r_{gi} \tag{5.22}$$

When noise is included we get, rather, the result (5.20), which can be written,

$$\left\{r_{ij} + \eta\, I_{ij}\right\} p_j = r_{gi} \qquad \text{or} \qquad p_j = \left\{r_{ij} + \eta\, I_{ij}\right\}^{-1} r_{gi} \tag{5.23}$$

Where I_{ij} is the unit matrix. Looking at the right-hand member of the pair of equations (5.23), it is easy to see that the coefficients p_j will be smaller when the error is large. This is most obvious if we assume that r_{ij} is diagonal. Thus we see that the inclusion of error makes the coefficients in (5.11) smaller and that therefore, by (5.10), the estimate we make will be closer to climatology. If we include error, then *OI* will draw more closely to climatology or the first guess and tend to weight new observations less heavily. This is desirable. By putting different values of η_j along the diagonal, one can put information on the confidence one has in individual stations into the analysis scheme and weight more heavily those stations in which one has more confidence.

What do we need to make *OI* work?

In order to make the above schemes work, we need the correlation matrices r_{ij} and r_{gi}. The first of these is easily calculable from observations, but the second is not since it involves correlations between the station points and the grid points. We do not have data at the grid points, or we wouldn't need an analysis scheme. In practice, not even the r_{ij} are calculated in full generality. It is possible to assume that correlations between points depend only on the distance between them and not on location or direction (although it would be possible to include directionally dependent (anisotropic) correlations). In this case the single isotropic correlation function can be estimated from station data. This is a crude approximation since correlations between stations depend on the location of the stations and whether longitude or latitude separates them. Some examples illustrating the anisotropy of correlation functions are shown on the following pages (Figs 5.3 and 5.4).

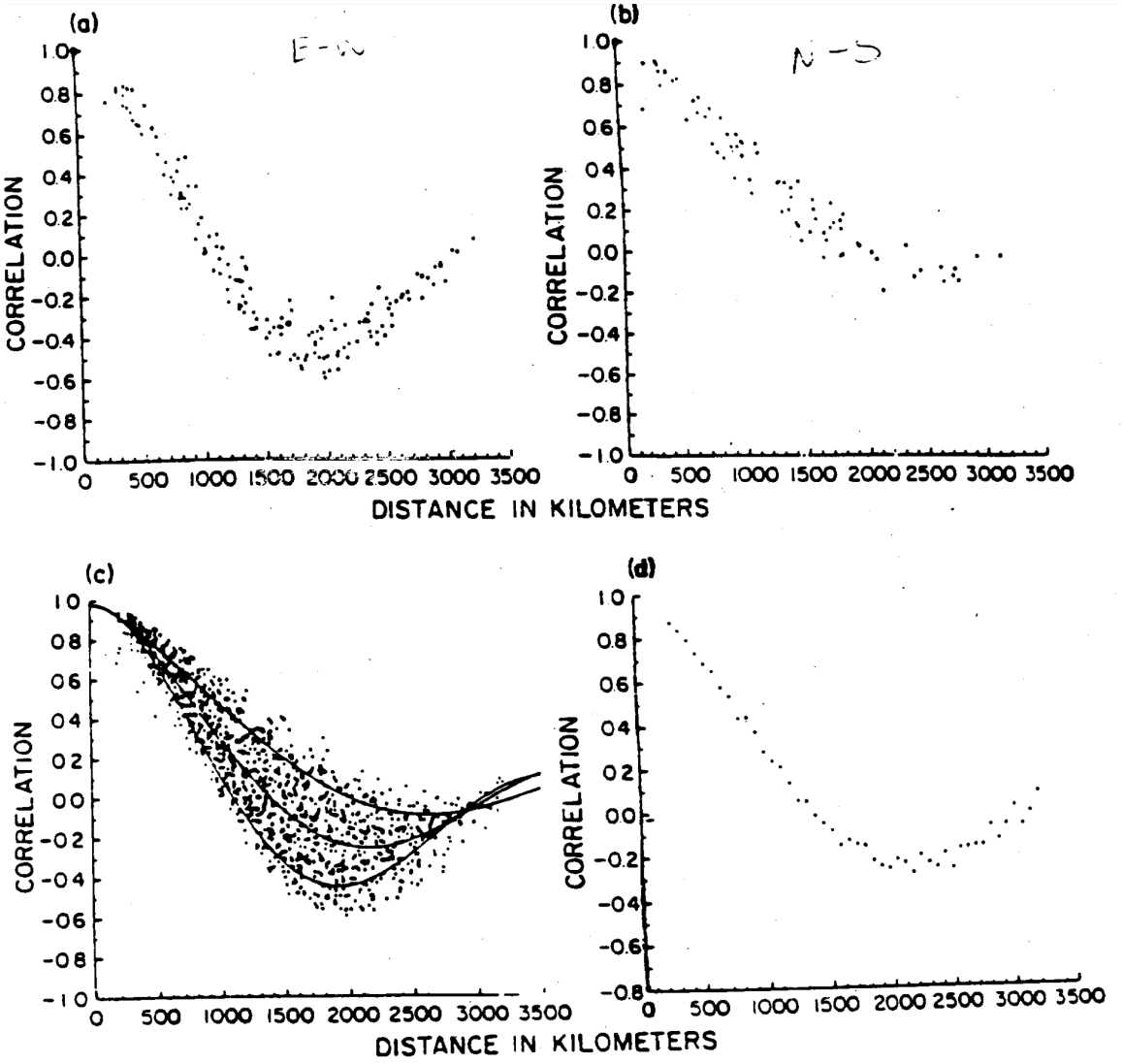


Fig. 5.3 . Example of anisotropy of geopotential correlation using 500 mb winter data versus between-station distance. a) Height-height correlations for East-West orientation, b) North-South orientation, c) Full array of correlations with top curve fitted to North-South variations and bottom curve fitted to East-West correlations, middle curve to whole array of correlations. D) average correlations values for 50 successive distance intervals of 62.5km. After H.J.Thiebeaux.

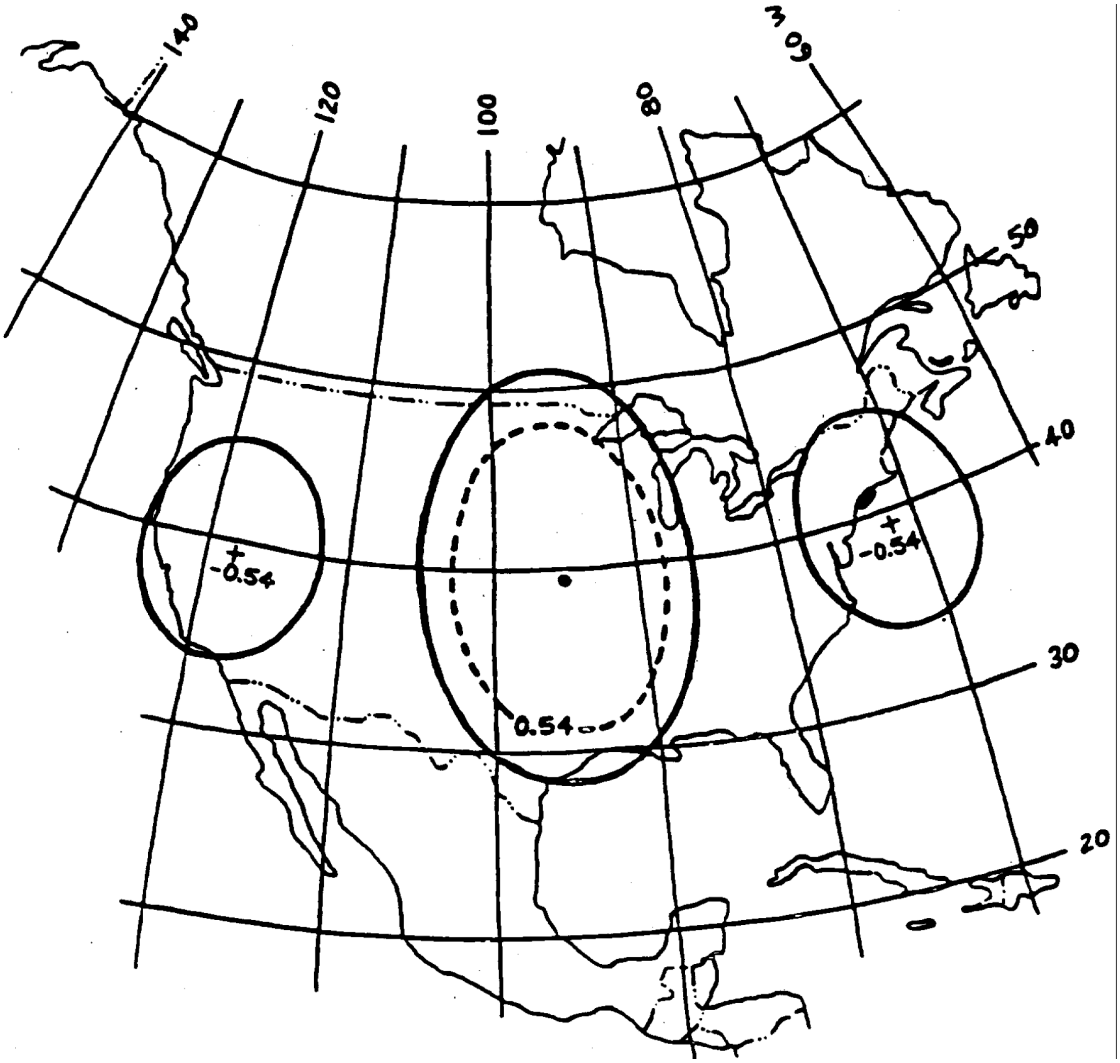


Figure 5.4. Anisotroic correlation contours, relative to Topeka, Kansas, reated by the two-dimensional autoregressive correlation model. Solid line ellipses are contours on which the 500mb geopotential correlations with Topeka have magnitude 0.35. Dashed line ellise and +'s are loci of correlation magnitude 0.54. After H.J. Thiebeaux.

5.4 Practical Optimal Interpolation of Meteorological Data

Suppose we wish to produce an optimal map of wind and geopotential on a pressure surface defined on a regular grid, given a bunch of scattered observations at station points. It is also possible to consider a multi-level or three-dimensional optimal interpolation, but it is just a messy extension of the example we will give here. This example follows Schlatter (1975). The vector of state variables at a grid point is defined by $\mathbf{z}$. It is expressed in terms of the first guess of the vector at the grid point, $\mathbf{z}_f$, corrected by an optimal weighting of the deviations of the observations vector at N nearby station points, $\mathbf{x}_i$, from the first guess field at those station points, $\mathbf{x}_{fi}$. The observation vectors at both the grid point and the surrounding station points consist of two wind components and height of the pressure surface. We write this as,

$$\begin{Bmatrix} u \\ v \\ h \end{Bmatrix} = \mathbf{z} = \mathbf{z}_f + \sum_{i=1}^{n} A_i \left(\mathbf{x}_i - \mathbf{x}_{fi} \right) \tag{5.24}$$

Note that the correction has been expressed in terms of a deviation of the observation at the station point from a first guess based on the previous analysis or some other *a priori* information. This is weighted by the coefficient A, with subscript to indicate the station. Each A_i may indeed be a 3x3 matrix, which converts the vector of deviations at each station point into a vector of corrections at the grid point. We can bring the summation into the matrix operation by extending the deviation vector to include all n station points, so that $\mathbf{x}$ and $\mathbf{x}_f$ are each 3n long. Then $\mathbf{A}$ becomes a 3x3n matrix. In this case, we can write the matrix problem for finding the best guess at the grid point, $\mathbf{z}$, as,

$$\mathbf{z} = \mathbf{z}_f + \mathbf{A} \left(\mathbf{x} - \mathbf{x}_f \right) \tag{5.25}$$

So the vector $\mathbf{z}$ of length 3 is determined by adding a correction to the first guess, $\mathbf{z}f$, which is determined by multiplying the 3x3n matrix $\mathbf{A}$ times the vector of deviations at the station points $\mathbf{x}$-$\mathbf{x}_f$, which is 3n long. The weighting coefficients in the matrix are determined by minimizing the inner product of the error vector, which is the difference between the estimated state vector at the grid point, $\mathbf{z}$, and its true value $\mathbf{z}_t$.

$$\left\langle \left(\mathbf{z}_t - \mathbf{z} \right) \left(\mathbf{z}_t - \mathbf{z} \right)^T \right\rangle \tag{5.26}$$

where $\langle\ \rangle$ represents an average over an ensemble of cases, or the expected value operator, if you like. We can substitute (5.25) into (5.26) and perform the minimization with respect to the matrix of coefficients $\mathbf{A}$.

$$\begin{aligned}\left\langle (\mathbf{z}_t - \mathbf{z})(\mathbf{z}_t - \mathbf{z})^T \right\rangle &= \left\langle \left(\mathbf{z}_t - \mathbf{z}_f - \mathbf{A}\left(\mathbf{x} - \mathbf{x}_f\right)\right)\left(\mathbf{z}_t - \mathbf{z}_f - \mathbf{A}\left(\mathbf{x} - \mathbf{x}_f\right)\right)^T \right\rangle \\ &= \left\langle \mathbf{z}_t - \mathbf{z}_f \right\rangle \left\langle \mathbf{z}_t - \mathbf{z}_f \right\rangle^T - \left\langle \mathbf{z}_t - \mathbf{z}_f \right\rangle \left\langle \mathbf{A}\left\langle \mathbf{x} - \mathbf{x}_f \right\rangle \right\rangle^T \\ &\quad - \mathbf{A}\left\langle \mathbf{x} - \mathbf{x}_f \right\rangle \left\langle \mathbf{z}_t - \mathbf{z}_f \right\rangle^T + \mathbf{A}\left\langle \mathbf{x} - \mathbf{x}_f \right\rangle \left(\mathbf{A}\left\langle \mathbf{x} - \mathbf{x}_f \right\rangle\right)^T\end{aligned} \tag{5.27}$$

Next we differentiate (5.27) with respect to the matrix of coefficients $\mathbf{A}$, you can show by trying to do the problem in scalar form that differentiation with respect to matrices is similar to differentiation with respect to scalars. If we perform the differentiation of (5.27) with $\mathbf{A}$ and set the result to zero, we obtain the desired expression for the coefficient matrix.

$$\begin{aligned}0 = &-\left\langle \mathbf{z}_t - \mathbf{z}_f \right\rangle \left\langle \mathbf{x} - \mathbf{x}_f \right\rangle^T - \left\langle \mathbf{x} - \mathbf{x}_f \right\rangle \left\langle \mathbf{z}_t - \mathbf{z}_f \right\rangle^T \\ &+ \left\langle \mathbf{x} - \mathbf{x}_f \right\rangle \left(\mathbf{A}\left\langle \mathbf{x} - \mathbf{x}_f \right\rangle\right)^T + \mathbf{A}\left\langle \mathbf{x} - \mathbf{x}_f \right\rangle \left\langle \mathbf{x} - \mathbf{x}_f \right\rangle^T\end{aligned}$$

After a little rearrangement and noting that, if a matrix product is zero, then its transpose will be zero also, we find the condition that $\mathbf{A}$ must satisfy,

$$0 = -\left\langle \mathbf{z}_t - \mathbf{z}_f \right\rangle \left\langle \mathbf{x} - \mathbf{x}_f \right\rangle^T + \mathbf{A}\left\langle \mathbf{x} - \mathbf{x}_f \right\rangle \left\langle \mathbf{x}_t - \mathbf{x}_f \right\rangle^T \tag{5.28}$$

which can be solved for the matrix of coefficients,

$$\mathbf{A} = \left\langle \mathbf{z}_t - \mathbf{z}_f \right\rangle \left\langle \boldsymbol{x} - \boldsymbol{x}_f \right\rangle^T \left(\left\langle \boldsymbol{x} - \boldsymbol{x}_f \right\rangle \left\langle \boldsymbol{x} - \boldsymbol{x}_f \right\rangle^T \right)^{-1} \tag{5.29}$$

or

$$\mathbf{A} = \underline{\mathbf{C}}_{\boldsymbol{zx}} \left(\underline{\mathbf{C}}_{\boldsymbol{xx}} \right)^{-1}$$

where $\underline{\mathbf{C}}_{\boldsymbol{z}\,\boldsymbol{x}}$ is the covariance matrix between the grid point and the station points and $\underline{\mathbf{C}}_{\boldsymbol{xx}}$ is the covariance matrix of the station points. The $\mathbf{A}$ matrix has nxJ^2 coefficients. If

J=3 variables and n=5 stations are used to define each grid value, then the matrix has 45 coefficients for each grid point. $\underline{\mathbf{C}}_{xx}$ is readily calculable from observations, since it is just the covariance matrix between observations. Rather than specify a different matrix for every set of stations, we might make $\underline{\mathbf{C}}_{xx}$ a function only of distance between the points and latitude, for example. $\underline{\mathbf{C}}_{zx}$ cannot be calculated directly, since we don't know the values at the grid points. We can assume again that the covariance is only a function of distance, and obtain $\underline{\mathbf{C}}_{zx}$ from $\underline{\mathbf{C}}_{xx}$. This is a lot like Cressman's correction scheme, except that the influence function is calculated from the behavior of the field itself, and the observed statistical relationships between winds and velocities are used (they are contained in the covariance matrices). With a little more work you can also objectively take account of the uncertainty of the measurements in an optimum way.

5.5 Operational Data Assimilation for Weather Forecasting

In weather forecasting, the weather prediction model itself is used to determine the best analysis of available observations with which to begin a forecast. Often, the data are modified to suppress gravity waves (normal mode initialization) that would otherwise adversely affect the forecast. These initialized data, which are heavily influenced by the model dynamics, physics and numeric are then used in diagnostic studies of weather and climate. Variational techniques are used, with the current state of the art being four-dimensional variational assimilation of data. In Chapter 8 of these notes we give a general introduction to the Kalman Filter, which is believed to be the ultimate method of optimally incorporating observations into a numerical model to produce the best analysis of an evolving dynamical system.

In discussing the data assimilation problem for weather forecasting, one must get specifically into the dynamical equations of atmospheric flow, which would take us beyond the focus of this course. Data assimilation is discussed more fully in Daley(1991), Kalnay(2002), Eversen(2009) and Park(2009). You will find that in this course we cover the material in about the first half of Daley's book. Least squares analysis of the type we have been discussing forms a big part of modern operational data assimilation. The Kalman Filter is discussed near the end of Daley's book and in greater modern detail by Evensen(2009). It is also beginning to be used in oceanography (Ballabrera, et al. 2001) and in hydrology (Reichle, et al., 2002).

5.6 Data Reconstruction Methods

In some cases, such as sea surface temperatures and other surface measurements, modern data sets are able to map global fields accurately using optimum interpolation (OI) or other techniques. As one goes back earlier in time, however, insufficient spatial coverage is available to accurately construct global maps. In such cases people have used reconstruction methods that use the known structure from modern observations to

develop structure functions that can then be used to interpolate sparse data from earlier periods. An example of this is Smith et al (1996) who used such methods to improve the analysis of SST prior to 1981. Data after 1981 include satellite estimates of SST, which give nearly global data at frequent intervals, and these data were used to construct EOFs that then were used to interpolate the sparser data from 1950-1981.

Suppose we write the SST in terms of a mean value and an EOF reconstruction,

$$T(x,t) = \overline{T}(x) + \mathbf{E}(x)\mathbf{Z}(t) \tag{5.30}$$

Where we will assume that the time mean remains constant with time, but we could take its variability into account easily enough, if we knew it. $\mathbf{E}(x)\mathbf{Z}(t)$ is an EOF expansion of the variability of the SST around the long-term mean $\overline{T}(x)$. The spatial structure in latitude and longitude is represent by a single spatial variable x.

Now suppose that we have a period in which we have observations of SST at a subset of the total points, K, and we will use a truncated expansion in EOFs choosing to use a subset M < m of the total possible EOFs at our disposal. We will write the anomalies $T'(x,t) = T(x,t) - \overline{T}(x)$ at those previous times as a truncated expansion of the EOFs, which are truncated at M and evaluated only at the points where we have data in the earlier time period.

$$\hat{T}(x,t) = \breve{\mathbf{E}}(x)\mathbf{W}(t) \tag{5.31}$$

Where $\breve{\mathbf{E}}(x)$ is a truncated EOF matrix including only the first M eigenvectors, and $\mathbf{W}(t)$ is a set of unknown weights to give the best fit at all x points, from observations at a subset of points available in the old record.

Next, of course, we need to define an error functional that we will minimize to determine the weights.

$$\varepsilon^2(t) = \sum_{i-1}^{K}\left(T(x_i,t) - \hat{T}(x_i,t)\right)^2 = \sum_{i-1}^{K}\left(T(x_i,t) - \breve{\mathbf{E}}(x_i)\mathbf{W}(t)\right)^2 \tag{5.32}$$

where the summation is over only the spatial grid points where we have data in the older record. Ideally, K > M so that the data will allow us to determine the weights for M eigenvectors. If we redefine the Temperature and EOF matrices as being defined only at the x points where we now have data, then we can write the solution as

$$\breve{\mathbf{E}}\mathbf{W}\breve{\mathbf{E}}^T = \breve{\mathbf{E}}\mathbf{T}^T \tag{5.33}$$

Where the derivation goes very much like that in section (3.4.1). This is a set of linear equations for the coefficients W. If the inverses of $\breve{\mathbf{E}}$ and $\breve{\mathbf{E}}^T$ are well behaved, then we can write the solution for the weights as, $\mathbf{W} = \mathbf{T}^T \breve{\mathbf{E}}^{T-1}$. So at each time step, we can stamp out the weighting coefficients from the known structure of $\breve{\mathbf{E}}^{T-1}$. Note that if we had data at all the original data points, $\breve{\mathbf{E}}^{T-1} = E$, and $W = \mathbf{T}^T E = Z$, the principal components.

I tried this, and to invert the truncated E with values only at the sample points you have to make it square, by making the number of retained eigenvectors equal to the number of sample points. It then works nearly perfectly at the sample points, but when you apply it to all the data points it blows up between the sample points like any overfitted functional fit. So what you have to do is make the number of EOFs retained fewer than the number of sample points. Then you can do a pseudo-inverse as in Chapter 3, or just a straight ahead best regression using some other method.

References: Objective Gridding of Data

Ballabrera, P.-J., A. J. Busalacchi, and R. Murtugudde, 2001, Application of a reduced-order Kalman filter to initialize a coupled atmosphere-ocean model: impact on the prediction of El Nino: Journal of Climate, **14**, 1720-37.

Berthorsson, P., and B. Doos, 1955: Numerical weather map analysis. *Tellus*, **7**, 329-340.

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

Daley, Roger, 1991: *Atmospheric Data Analysis*. Cambridge Univ. Press, 457 pp.

Evensen, Geir, 2009: Data Assimilation: the Ensemble Kalman Filter, Springer, 330pp.

Flattery, T.W., 1971: Spectral models for global analysis and forecasting. Proc. AWS Tech. Exch. Conf. U.S. Naval Academy, Sept. 1970, 42-54. (See also Haltiner and Williams, 1980, pp. 377-379.)

Gandin, L.S., 1963: *Objective Analysis of Meteorological Fields*, Gidrometeorologicheskoe Izdatel'stvo, Leningrad, Translated from Russian, Israel Program for Scientific Translations, Jerusalem, 1965, 242 pp.

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

Haltiner, G.J., 1970: *Numerical Weather Prediction*. Wiley, 317pp.

Haltiner, G.J., and R. T. Williams, 1980: *Numerical Prediction and Dynamic Meteorology*, Second Ed. Wiley, 477 pp.

Kalney, Eugenia, 2002: Atmospheric Modeling, Data Assimilation and Predictability, Cambridge University Press, 364pp.

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

Park, Seon K., Ed., 2009: Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications, Springer, 493pp.

Reichle, R. H., D. B. McLaughlin, and D. Entekhabi, 2002, Hydrologic Data Assimilation with the Ensemble Kalman Filter: Mon. Wea. Rev., **130**, 103–114.

Schlatter, T.W., 1975: Some experiments with a multivariate statistical objective analysis scheme. *Mon. Wea. Rev.*, **103**, 246-257.

Schlatter, T.W., G.W. Branstator, and L.G. Thiel, 1976: Testing a global multivariate statistical objective analysis scheme with observed data. *Mon.Wea. Rev.*, **104**, 765-783.

Smith, T. M., R. W. Reynolds, et al. (1996). "Reconstruction of historical sea surface temperatures using empirical orthogonal functions." J. Climate **9**(6): 1403-1420.

Thiébaux, H.J., and M.A. Pedder, 1987: *Spatial Objective Analysis with Applications in Atmospheric Science*. Academic Press, 299 pp.

Wiener, N. 1949: *The Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications*.Wiley, New York.