Chapter 7

Statistical Forecasting

7.1. BACKGROUND

Much of operational weather and long-range (seasonal, or “climate”) forecasting has a statistical basis. As a nonlinear dynamical system, the atmosphere is not perfectly predictable in a deterministic sense. Consequently, statistical methods are useful, and indeed necessary, parts of the forecasting enterprise. This chapter provides an introduction to statistical forecasting of scalar (single-number) quantities. Some methods suited to statistical prediction of vector (multiple values simultaneously) quantities, for example, spatial patterns, are presented in Sections 13.2.3 and 14.4.

Some statistical forecast methods operate without information from the fluid-dynamical forecast models that have become the mainstay of weather forecasting for lead times ranging from one day to a week or so in advance. Such pure statistical forecast methods are sometimes referred to as Classical, reflecting their prominence in the years before dynamical forecast information was available. These methods continue to be viable and useful at very short lead times (hours in advance), or very long lead times (weeks or more in advance), for which the dynamical forecast information is not available with sufficient promptness or accuracy, respectively.

Another important application of statistical methods to weather forecasting is in conjunction with dynamical forecast information. Statistical forecast equations routinely are used to postprocess and enhance the results of dynamical forecasts at operational weather forecasting centers throughout the world, and are essential as guidance products to aid weather forecasters. The combined statistical and dynamical approaches are especially important for providing forecasts for quantities and locations (e.g., particular cities rather than gridpoints) not represented by the dynamical models.

The types of statistical forecasts mentioned so far are objective, in the sense that a given set of inputs always produces the same particular output. However, another important aspect of statistical weather forecasting is in the subjective formulation of forecasts, particularly when the forecast quantity is a probability or set of probabilities. Here the Bayesian interpretation of probability as a quantified degree of belief is fundamental. Subjective probability assessment forms the basis of many operationally important forecasts and is a technique that could be used more broadly to enhance the information content of operational forecasts.

7.2. LINEAR REGRESSION

Much of statistical weather forecasting is based on the procedure known as linear, least-squares regression. In this section, the fundamentals of linear regression are reviewed. Much more complete treatments can be found in standard texts such as Draper and Smith (1998) and Neter et al. (1996).

Statistical Methods in the Atmospheric Sciences. DOI: 10.1016/B978-0-12-385022-5.00007-5

7.2.1. Simple Linear Regression

Regression is most easily understood in the case of *simple linear regression*, which describes the linear relationship between two variables, say x and y. Conventionally, the symbol x is used for the *independent*, or *predictor variable*, and the symbol y is used for the *dependent variable*, or *predictand*. More than one predictor variable is very often required in practical forecast problems, but the ideas for simple linear regression generalize easily to this more complex case of *multiple linear regression*. Therefore, most of the important ideas about regression can be presented in the context of simple linear regression.

Essentially, simple linear regression seeks to summarize the relationship between x and y, shown graphically in their scatterplot, using a single straight line. The regression procedure chooses that line producing the least error for predictions of y given observations of x. Exactly what constitutes least error can be open to interpretation, but the most usual error criterion is minimization of the sum (or, equivalently, the average) of the squared errors. It is the choice of the squared-error criterion that is the basis of the name *least-squares regression*, or *ordinary least squares* (OLS) regression. Other error measures are possible, for example, minimizing the average (or, equivalently, the sum) of absolute errors, which is known as *least absolute deviation* (LAD) *regression* (Gray et al,. 1992; Mielke et al., 1996). Choosing the squared-error criterion is conventional not because it is necessarily best, but rather because it makes the mathematics analytically tractable. Adopting the squared-error criterion results in the line-fitting procedure being fairly tolerant of small discrepancies between the line and the points. However, the fitted line will adjust substantially to avoid very large discrepancies. It is thus not resistant to outliers. Alternatively, LAD regression is resistant because the errors are not squared, but the lack of analytic results (formulas) for the regression function means that the estimation must be iterative.

Figure 7.1 illustrates the situation. Given a data set of (x, y) pairs, the problem is to find the particular straight line,

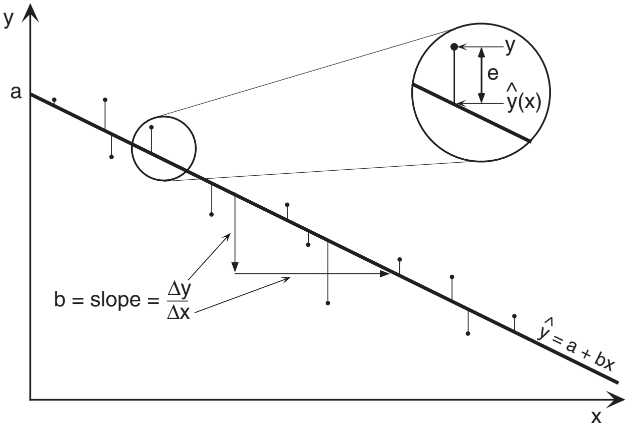


FIGURE 7.1 Schematic illustration of simple linear regression. The regression line, $\hat{y} = a + bx$, is chosen as the one minimizing some measure of the vertical differences (the residuals) between the points and the line. In least-squares regression that measure is the sum of the squared vertical distances. The inset shows a residual, e, as the difference between a data point and the regression line.

$$\hat{y} = a + bx, \tag{7.1}$$

minimizing the squared vertical distances (thin lines) between it and the data points. The circumflex ("hat") accent signifies that the equation specifies a predicted value of y. The inset in Figure 7.1 indicates that the vertical distances between the data points and the line, also called errors or *residuals*, are defined as

$$e_i = y_i - \hat{y}(x_i). \tag{7.2}$$

There is a separate residual e_i for each data pair (x_i, y_i). Note that the sign convention implied by Equation 7.2 is for points above the line to be regarded as positive errors, and points below the line to be negative errors. This is the usual convention in statistics, but is opposite to what often is seen in the atmospheric sciences, where forecasts smaller than the observations (the line being below the point) are regarded as having negative errors, and vice versa. However, the sign convention for the residuals is unimportant, since it is the minimization of the sum of squared residuals that defines the best-fitting line. Combining Equations 7.1 and 7.2 yields the regression equation,

$$y_i = \hat{y}_i + e_i = a + bx_i + e_i, \tag{7.3}$$

which says that the true value of the predictand is the sum of the predicted value (Equation 7.1) and the residual.

Finding analytic expressions for the least-squares intercept, a, and the slope, b, is a straightforward exercise in calculus. In order to minimize the sum of squared residuals,

$$\sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - [a + bx_i])^2, \tag{7.4}$$

it is only necessary to set the derivatives of Equation 7.4 with respect to the parameters a and b to zero and solve. These derivatives are

$$\frac{\partial \sum_{i=1}^{n} (e_i)^2}{\partial a} = \frac{\partial \sum_{i=1}^{n} (y_i - a - bx_i)^2}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \tag{7.5a}$$

and

$$\frac{\partial \sum_{i=1}^{n} (e_i)^2}{\partial b} = \frac{\partial \sum_{i=1}^{n} (y_i - a - bx_i)^2}{\partial b} = -2\sum_{i=1}^{n} x_i[(y_i - a - bx_i)] = 0. \tag{7.5b}$$

Rearranging Equations 7.5 leads to the so-called *normal equations*,

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \tag{7.6a}$$

and

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} (x_i)^2. \tag{7.6b}$$

Dividing Equation 7.6a by n leads to the observation that the fitted regression line must pass through the point located by the two sample means of x and y. Finally, solving the normal equations for the regression parameters yields

$$b = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n}(x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{7.7a}$$

and

$$a = \bar{y} - b\bar{x}. \tag{7.7b}$$

Equation 7.7a, for the slope, is similar in form to the Pearson correlation coefficient and can be obtained with a single pass through the data using the computational form given as the second equality. Note that, as was the case for the correlation coefficient, careless use of the computational form of Equation 7.7a can lead to roundoff errors since the numerator may be the difference between two large numbers.

7.2.2. Distribution of the Residuals

Thus far, fitting the straight line has involved no statistical ideas at all. All that has been required was to define least error to mean minimum squared error. The rest has followed from straightforward mathematical manipulation of the data, namely, the (x, y) pairs. To bring in statistical ideas, it is conventional to assume that the quantities e_i are independent random variables with zero mean and constant variance. Often, the additional assumption is made that these residuals follow a Gaussian distribution.

Assuming that the residuals have zero mean is not at all problematic. In fact, one convenient property of the least-squares fitting procedure is the guarantee that

$$\sum_{i=1}^{n} e_i = 0, \tag{7.8}$$

from which it is clear that the sample mean of the residuals (dividing this equation by n) is also zero.

Imagining that the residuals can be characterized in terms of a variance is really the point at which statistical ideas begin to come into the regression framework. Implicit in their possessing a variance is the idea that the residuals scatter randomly about some mean value (Equation 4.21 or 3.6). Equation 7.8 says that the mean value around which they will scatter is zero, so it is the regression line around which the data points will scatter. We then need to imagine a series of distributions of the residuals *conditional* on the x values, with each observed residual regarded as having been drawn from one of these conditional distributions. The constant variance assumption really means that the variance of the residuals is constant in x, or that all of these conditional distributions of the residuals have the same variance. Therefore a given residual (positive or negative, large or small) is by assumption equally likely to occur at any part of the regression line.

Figure 7.2 is a schematic illustration of the idea of a suite of conditional distributions centered on the regression line. The three small gray distributions are identical, except that their means are shifted higher or lower depending on the level of the regression line (predicted value of y) for each x.

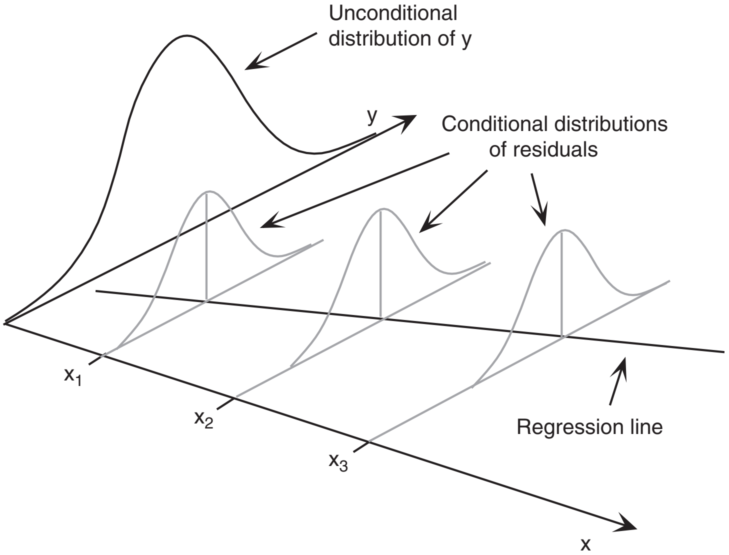

Unconditional distribution of y
y
Conditional distributions of residuals
x1
x2
x3
Regression line
x

FIGURE 7.2 Schematic illustration of distributions (gray) of residuals around the regression line, conditional on these values of the predictor variable, *x*. The actual residuals are regarded as having been drawn from these distributions.

Extending this thinking slightly, it is not difficult to see that the regression equation can be regarded as specifying the conditional mean of the predictand, given a specific value of the predictor. Also shown by the large black distribution in Figure 7.2 is a schematic representation of the unconditional distribution of the predictand, *y*. The distributions of residuals are less spread out (have smaller variance) than the unconditional distribution of *y*, indicating that there is less uncertainty about *y* if a corresponding *x* value is known.

Central to the making of statistical inferences in the regression setting is estimation of this (constant) residual variance from the sample of residuals. Since the sample average of the residuals is guaranteed by Equation 7.8 to be zero, the square of Equation 3.6 becomes

$$s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2, \tag{7.9}$$

where the sum of squared residuals is divided by *n* – 2 because two parameters (*a* and *b*) have been estimated. Substituting Equation 7.2 then yields

$$s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n} [y_i - \hat{y}(x_i)]^2. \tag{7.10}$$

Rather than compute the estimated residual variance using 7.10, however, it is more usual to use a computational form based on the relationship,

$$SST = SSR + SSE, \tag{7.11}$$

which is proved in most regression texts. The notation in Equation 7.11 consists of acronyms describing the variation in the predictand, *y* (SST), and a partitioning of that variation between the portion represented by the regression (SSR), and the unrepresented portion ascribed to the variation of the residuals (SSE). The term SST is an acronym for sum of squares, total, which has the mathematical meaning of the sum of squared deviations of the y values around their mean,

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2. \tag{7.12}$$

This term is proportional (by the factor $n - 1$) to the sample variance of y and thus measures the overall variability of the predictand. The term SSR stands for the regression sum of squares, or the sum of squared differences between the regression predictions and the sample mean of y,

$$SSR = \sum_{i=1}^{n} [\hat{y}(x_i) - \bar{y}]^2, \tag{7.13a}$$

which relates to the regression equation according to

$$SSR = b^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = b^2 \left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right]. \tag{7.13b}$$

Equation 7.13 indicates that a regression line differing little from the sample mean of the y values will have a small slope and produce a very small SSR, whereas one with a large slope will exhibit some large differences from the sample mean of the predictand and therefore produce a large SSR.

Finally, SSE refers to the sum of squared errors, or sum of squared differences between the residuals and their mean, which is zero,

$$SSE = \sum_{i=1}^{n} e_i^2. \tag{7.14}$$

Since this differs from Equation 7.9 only by the factor of n–2, rearranging Equation 7.11 yields the computational form

$$s_e^2 = \frac{1}{n-2}\{SST - SSR\} = \frac{1}{n-2}\left\{ \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - b^2 \left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right] \right\}. \tag{7.15}$$

7.2.3. The Analysis of Variance Table

In practice, regression analysis is now almost universally done using computer software. A central part of the regression output from these software packages is a summary of the foregoing information in an *analysis of variance*, or ANOVA table. Usually, not all the information in an ANOVA table will be of interest, but it is such a universal form of regression output that you should understand its components. Table 7.1 outlines the arrangement of an ANOVA table for simple linear regression and indicates where the quantities described in the previous section are reported. The three rows correspond to the partition of the variation of the predictand as expressed in Equation 7.11. Accordingly, the Regression and Residual entries in the df (degrees of freedom) and SS (sum of squares) columns will sum to the corresponding entry in the Total row. Therefore, the ANOVA table contains some redundant information, and as a consequence the output from some regression packages will omit the Total row entirely.

TABLE 7.1 Generic analysis of variance (ANOVA) table for simple linear regression. The column headings df, SS, and MS stand for degrees of freedom, sum of squares, and mean square, respectively. Regression df = 1 is particular to simple linear regression (i.e., a single predictor x). Parenthetical references are to equation numbers in the text.

Source	df	SS	MS	F
Total	$n - 1$	SST (7.12)		
Regression	1	SSR (7.13)	MSR = SSR / 1	(F = MSR/MSE)
Residual	$n - 2$	SSE (7.14)	MSE = s_e^2	

The entries in the MS (mean squared) column are given by the corresponding quotients of SS/df. For simple linear regression, the regression df = 1, and SSR = MSR. Comparing with Equation 7.15, it can be seen that the MSE (mean squared error) is the estimated sample variance of the residuals. The total mean square, left blank in Table 7.1 and in the output of most regression packages, would be SST/(n – 1), or simply the sample variance of the predictand.

7.2.4. Goodness-of-Fit Measures

The ANOVA table also presents (or provides sufficient information to compute) three related measures of the fit of a regression, or the correspondence between the regression line and a scatterplot of the data. The first of these measures is the MSE. From the standpoint of forecasting, the MSE is perhaps the most fundamental of the three measures, since it indicates the variability of, or the uncertainty about, the observed y values (the quantities being forecast) around the forecast regression line. As such, it directly reflects the average accuracy of the resulting forecasts. Referring again to Figure 7.2, since MSE = s_e^2 this quantity indicates the degree to which the distributions of residuals cluster tightly (small MSE) or spread widely (large MSE) around a regression line. In the limit of a perfect linear relationship between x and y, the regression line coincides exactly with all the point pairs, the residuals are all zero, SST will equal SSR, SSE will be zero, and the variance of the residual distributions is also zero. In the opposite limit of absolutely no linear relationship between x and y, the regression slope will be zero, the SSR will be zero, SSE will equal SST, and the MSE will very nearly equal the sample variance of the predictand itself. In this unfortunate case, the three conditional distributions in Figure 7.2 would be indistinguishable from the unconditional distribution of y.

The relationship of the MSE to the strength of the regression fit is also illustrated in Figure 7.3. Panel (a) shows the case of a reasonably good regression, with the scatter of points around the regression line being fairly small. Here SSR and SST are nearly the same. Panel (b) shows an essentially useless regression, for values of the predictand spanning the same range as in panel (a). In this case the SSR is nearly zero since the regression has nearly zero slope, and the MSE is essentially the same as the sample variance of the y values themselves.

The second usual measure of the fit of a regression is the *coefficient of determination*, or R^2. This can be computed from

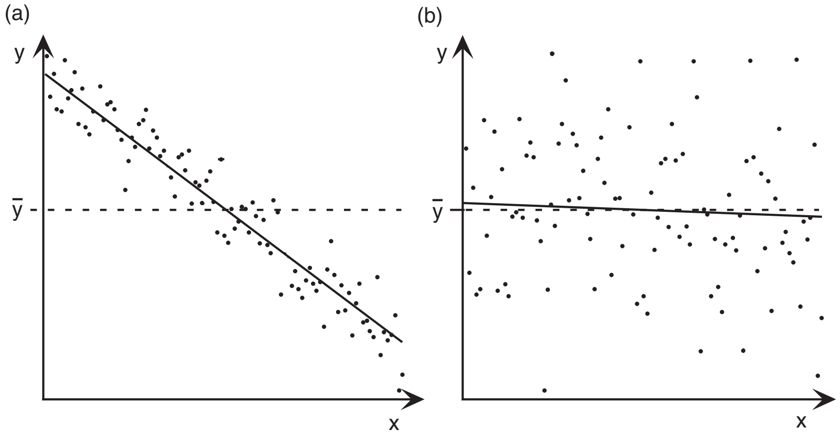


FIGURE 7.3 Illustration of the distinction between a fairly good regression relationship (a) and an essentially useless relationship (b). The points in panel (a) cluster closely around the regression line (solid), indicating small MSE, and the line deviates strongly from the average value of the predictand (dashed), producing a large SSR. In panel (b) the scatter around the regression line is large, and the regression line is almost indistinguishable from the mean of the predictand.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}, \tag{7.16}$$

which is often also displayed as part of standard regression output. The SSR is nearly equal to SST if each predicted value is close to its respective y, so that the corresponding residual is near zero. Therefore MSE and R^2 are different but related ways of expressing the closeness of or discrepancy between SST and SSR. The R^2 can be interpreted as the proportion of the variation of the predictand (proportional to SST) that is described or accounted for by the regression (SSR). Sometimes we see this concept expressed as the proportion of variation “explained,” although this claim is misleading: a regression analysis can quantify the nature and strength of a relationship between two variables but can say nothing about which variable (if either) causes the other. This is the same caveat offered in the discussion of the correlation coefficient in Chapter 3. For the case of simple linear regression, the square root of the coefficient of determination is exactly (the absolute value of) the Pearson correlation between x and y.

For a perfect regression, SSR = SST and SSE = 0, so $R^2 = 1$. For a completely useless regression, SSR = 0 and SSE = SST, so that $R^2 = 0$. Again, Figure 7.3b shows something close to this latter case. Comparing Equation 7.13a, the least-squares regression line is almost indistinguishable from the sample mean of the predictand, so SSR is very small. In other words, little of the variation in y can be ascribed to the regression, so the proportion SSR/SST is nearly zero.

The third commonly used measure of the strength of the regression is the F ratio, generally given in the last column of the ANOVA table. The ratio MSR/MSE increases with the strength of the regression, since a strong relationship between x and y will produce a large MSR and a small MSE. Assuming that the residuals are independent and follow the same Gaussian distribution, and under the null hypothesis of no real linear relationship, the sampling distribution of the F ratio has a known parametric form. This distribution forms the basis of a test that is applicable in the case of simple linear regression if the correct single predictor is known in advance of the analysis, but in the more general case of multiple regression (more than one x variable) problems of test multiplicity, to be discussed later, usually invalidate it. However, even if the F ratio cannot be used for quantitative statistical inference, it is still a valid qualitative index of the strength of a regression. See, for example, Draper and Smith (1998) or Neter et al. (1996) for discussions of the F test for overall significance of the regression.

7.2.5. Sampling Distributions of the Regression Coefficients

Another important use of the estimated residual variance is to obtain estimates of the sampling distributions of the regression coefficients. As statistics computed from a finite set of data subject to sampling variations, the computed regression intercept and slope, a and b, also exhibit sampling variability. That is, different batches of size n from the same data-generating process will yield different pairs of regression slopes and intercepts, and their sampling distributions characterize this batch-to-batch variability. Estimation of these sampling distributions allows construction of confidence intervals for the true population counterparts around the sample intercept and slope values a and b, and provides a basis for hypothesis tests about the corresponding population values.

Under the assumptions listed previously, the sampling distributions for both intercept and slope are Gaussian. On the strength of the Central Limit Theorem, this result also holds at least approximately for any regression when n is large enough because the estimated regression parameters (Equation 7.7) are obtained as the sums of large numbers of random variables. For the intercept the sampling distribution has parameters

$$\mu_a = a \tag{7.17a}$$

and

$$\sigma_a = s_e \left[\frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \right]^{1/2}. \tag{7.17b}$$

For the slope the parameters of the sampling distribution are

$$\mu_b = b \tag{7.18a}$$

and

$$\sigma_b = \frac{s_e}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{1/2}}. \tag{7.18b}$$

Equations 7.17a and 7.18a indicate that the least-squares regression parameter estimates are unbiased. Equations 7.17b and 7.18b show that the precision with which the intercept and slope can be estimated from the data depend directly on the estimated standard deviation of the residuals, s_e, which is the square root of the MSE from the ANOVA table (see Table 7.1). In addition, the estimated slope and intercept are not independent, having correlation

$$r_{a,b} = \frac{-\bar{x}}{\frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}}. \tag{7.19}$$

Taken together with the (at least approximately) Gaussian sampling distributions for a and b, Equations 7.17 through 7.19 define their joint bivariate normal (Equation 4.33) distribution. Equations 7.17b, 7.18b, and 7.19 are valid only for simple linear regression. With more than one predictor variable, analogous (vector) equations (Equation 10.40) must be used.

The output from regression packages will almost always include the standard errors (Equations 7.17b and 7.18b) in addition to the parameter estimates themselves. Some packages also include the ratios of the estimated parameters to their standard errors in a column labeled t ratio. When this is done, a one-sample t test (Equation 5.3) is implied, with the null hypothesis being that the underlying (population) mean for the parameter is zero. Sometimes a p value associated with this test is also automatically included in the regression output.

For the regression slope, this implicit t test bears directly on the meaningfulness of the fitted regression. If the estimated slope is small enough that its true value could plausibly (with respect to its sampling distribution) be zero, then the regression is not informative or useful for forecasting. If the slope is actually zero, then the value of the predictand specified by the regression equation is always the same and equal to its sample mean (cf. Equations 7.1 and 7.7b). If the assumptions regarding the regression residuals are satisfied, we would reject this null hypothesis at the 5% level if the estimated slope is, roughly, at least twice as large (in absolute value) as its standard error.

The same hypothesis test for the regression intercept often is offered by computerized statistical packages as well. Depending on the problem at hand, however, this test for the intercept may or may not be meaningful. Again, the t ratio is just the parameter estimate divided by its standard error, so the implicit null hypothesis is that the true intercept is zero. Occasionally, this null hypothesis is physically meaningful, and if so the test statistic for the intercept is worth looking at. On the other hand, it often happens that there is no physical reason to expect that the intercept might be zero. It may even be that a zero intercept is physically impossible. In such cases this portion of the automatically generated computer output is meaningless.

Example 7.1. A Simple Linear Regression

To concretely illustrate simple linear regression, consider the January 1987 minimum temperatures at Ithaca and Canandaigua from Table A.1 in Appendix A. Let the predictor variable, x, be the Ithaca minimum temperature, and the predictand, y, be the Canandaigua minimum temperature. The scatterplot of this data is shown in the middle panel of the bottom row of the scatterplot matrix in Figure 3.27 and as part of Figure 7.10. A fairly strong, positive, and reasonably linear relationship is indicated.

Table 7.2 shows what the output from a typical statistical computer package would look like for this regression. The data set is small enough that the computational formulas can be worked through to verify the results. (A little work with a hand calculator will verify that $\Sigma x = 403$, $\Sigma y = 627$, $\Sigma x^2 = 10803$, $\Sigma y^2 = 15009$, and $\Sigma xy = 11475$.) The upper portion of Table 7.2 corresponds to the template in Table 7.1, with the relevant numbers filled in. Of particular importance is MSE = 11.780, yielding as its square root the estimated sample standard deviation for the residuals, $s_e = 3.43°\text{F}$. This standard deviation addresses directly the precision of specifying the Canandaigua temperatures on the basis of the concurrent Ithaca temperatures, since we expect about 95% of the actual predictand values to be within $\pm 2s_e = \pm\ 6.9°\text{F}$ of the temperatures given by the regression. The coefficient of determination is easily computed as $R^2 = 1985.798/2327.419 = 85.3\%$. The Pearson correlation is $\surd 0.853 = 0.924$, as was given in Table 3.5. The value of the F statistic is very high, considering that the 99th percentile of its distribution under the null hypothesis of no real relationship is about 7.5. We also could compute

TABLE 7.2 Example output typical of that produced by computer statistical packages, for prediction of Canandaigua minimum temperature (*y*) using Ithaca minimum temperature (*x*) as the predictor, from the January 1987 data set in Table A.1.

Source	df	SS	MS	F
Total	30	2327.419		
Regression	1	1985.798	1985.798	168.57
Residual	29	341.622	11.780	

Variable	Coefficient	s.e.	t ratio
Constant	12.4595	0.8590	14.504
Ithaca Min	0.5974	0.0460	12.987

the sample variance of the predictand, which would be the total mean square cell of the table, as 2327.419/30 = 77.58°F^2.

The lower portion of Table 7.2 gives the regression parameters, *a* and *b*, their standard errors, and the ratios of these parameter estimates to their standard errors. The specific regression equation for this data set, corresponding to Equation 7.1, would be

$$T_{Can.} = \underset{(0.859)}{12.46} + \underset{(0.046)}{0.597}\ T_{Ith}. \tag{7.20}$$

Thus, the Canandaigua temperature would be estimated by multiplying the Ithaca temperature by 0.597 and adding 12.46°F. The intercept $a = 12.46$°F has no special physical significance except as the predicted Canandaigua temperature when the Ithaca temperature is 0°F. Notice that the standard errors of the two coefficients have been written parenthetically below the coefficients themselves. Although this is not a universal practice, it is very informative to someone reading Equation 7.20 without the benefit of the information in Table 7.2. In particular, it allows the reader to get a sense for the significance of the slope (i.e., the parameter *b*). Since the estimated slope is about 13 times larger than its standard error, it is almost certainly not really zero. This conclusion speaks directly to the question of the meaningfulness of the fitted regression. On the other hand, the corresponding implied hypothesis test for the intercept is much less interesting, because the possibility of a zero intercept has no physical significance. ◇

7.2.6. Examining Residuals

It is not sufficient to feed data to a computer regression package and uncritically accept the results. Some of the results can be misleading if the assumptions underlying the computations are not satisfied. Since these assumptions pertain to the residuals, it is important to examine the residuals for consistency with the assumptions made about their behavior.

One easy and fundamental check on the residuals can be made by examining a scatterplot of the residuals as a function of the predicted value $\hat{y}$. Many statistical computer packages provide this capability as a standard regression option. Figure 7.4a shows the scatterplot of a hypothetical data set, with the least-squares regression line, and Figure 7.4b shows a plot for the resulting residuals as a function of the predicted values. The residual plot presents the impression of "fanning," or exhibition of increasing spread as $\hat{y}$ increases. That is, the variance of the residuals appears to increase as the predicted value increases. This condition of nonconstant residual variance is called *heteroscedasticity*. Since the computer program that fit the regression has assumed constant residual variance, the MSE given in the ANOVA table is an overestimate for smaller values of x and y (where the points cluster closer to the regression line), and an underestimate of the residual variance for larger values of x and y (where the points tend to be further from the regression line). If the regression is used as a forecasting tool, we would be overconfident about forecasts for larger values of y and underconfident about forecasts for smaller values of y. In addition, the sampling distributions of the regression parameters will be more variable than implied by Equations 7.17 and 7.18. That is, the parameters will not have been estimated as precisely as the standard regression output would lead us to believe.

Often, nonconstancy of residual variance of the sort shown in Figure 7.4b can be remedied by transforming the predictand y, perhaps by using a power transformation (Equation 3.19 or 3.22). Figure 7.5 shows the regression and residual plots for the same data as in Figure 7.4 after logarithmically transforming the predictand. Recall that the logarithmic transformation reduces all the data values but reduces the

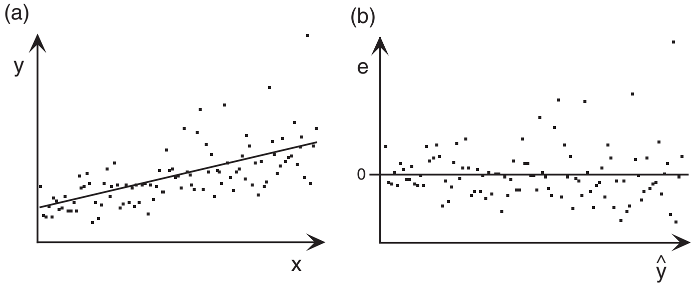


FIGURE 7.4 Hypothetical linear regression (a), and plot of the resulting residuals against the predicted values (b), for a case where the variance of the residuals is not constant. The scatter around the regression line in (a) increases for larger values of x and y, producing a visual impression of "fanning" in the residual plot (b). A transformation of the predictand is indicated.

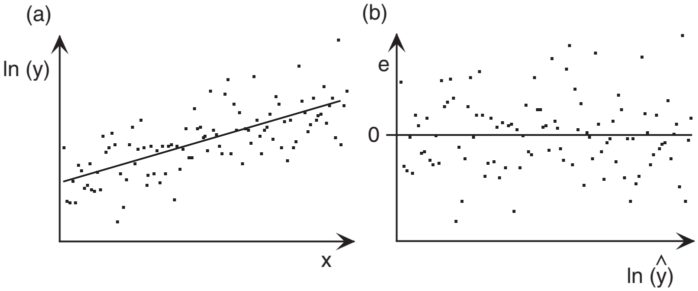


FIGURE 7.5 Scatterplots with regression (a) and resulting residual plot (b) for the same data in Figure 7.4, after logarithmically transforming the predictand. The visual impression of a horizontal band in the residual plot supports the assumption of constant variance of the residuals.

larger values more strongly than the smaller ones. Thus, the long right tail of the predictand has been pulled in relative to the shorter left tail, as in Figure 3.12. As a result, the transformed data points appear to cluster more evenly around the new regression line. Instead of fanning, the residual plot in Figure 7.5b gives the visual impression of a horizontal band, indicating appropriately constant variance of the residuals (*homoscedasticity*). Note that if the fanning in Figure 7.4b had been in the opposite sense, with greater residual variability for smaller values of $\hat{y}$ and lesser residual variability for larger values of $\hat{y}$, a transformation that stretches the right tail relative to the left tail (e.g., y^2) would have been appropriate.

It can also be informative to look at scatterplots of residuals as a function of a predictor variable. Figure 7.6 illustrates some of the forms such plots can take and their diagnostic interpretations. Figure 7.6a is similar to Figure 7.4b in that the fanning of the residuals indicates nonconstancy of variance. Figure 7.6b illustrates a different form of heteroscedasticity that might be more challenging to remedy through a variable transformation. The type of residual plot in Figure 7.6c, with a linear dependence on the predictor of the linear regression, indicates that either the intercept a has been omitted or that the calculations have been done incorrectly. Deliberately omitting a regression intercept, called “forcing through the origin,” is useful in some circumstances but may not be appropriate even if it is known beforehand that the true relationship should pass through the origin. Particularly if data are available over only a restricted range, or if the actual relationship is nonlinear, a linear regression including an intercept term may yield better predictions. In this latter case a simple linear regression would be similar to a first-order Taylor approximation about the mean of the training data.

Figure 7.6d shows a form for the residual plot that can occur when additional predictors would improve a regression relationship. Here the variance is reasonably constant in x, but the (conditional) average residual exhibits a dependence on x. Figure 7.6e illustrates the kind of behavior that can occur when a single outlier in the data has undue influence on the regression. Here the regression line has been pulled toward the outlying point in order to avoid the large squared error associated with it, leaving a trend in the other residuals. If the outlier were determined not to be a valid data point, it should either be corrected if possible or otherwise discarded. If it is a valid data point, a resistant approach

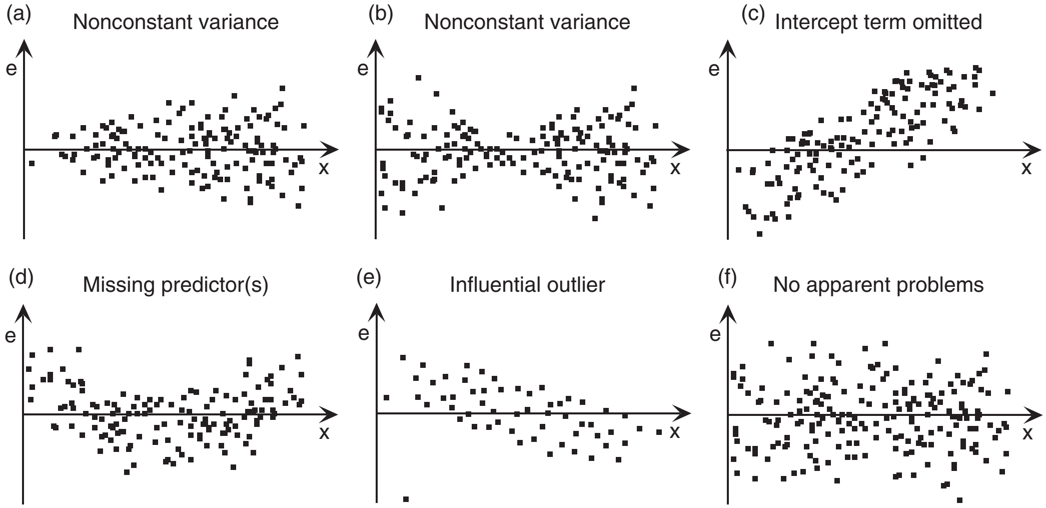


FIGURE 7.6 Idealized scatterplots of regression residuals versus a predictor x, with corresponding diagnostic interpretations.

such as LAD regression might be more appropriate. Figure 7.6f again illustrates the desirable horizontally banded pattern of residuals, similar to Figure 7.5b.

A graphical impression of whether the residuals follow a Gaussian distribution can be obtained through a Q–Q plot. Such plots are often a standard option in statistical computer packages. Figures 7.7a and 7.7b show Q–Q plots for the residuals in Figures 7.4b and 7.5b, respectively. The residuals are plotted on the vertical, and the standard Gaussian variables corresponding to the empirical cumulative probability of each residual are plotted on the horizontal. The curvature apparent in Figure 7.7a indicates that the residuals from the regression involving the untransformed predictand are positively skewed relative to the (symmetric) Gaussian distribution. The Q–Q plot of residuals from the regression involving the logarithmically transformed predictand is very nearly linear. Evidently the logarithmic transformation has produced residuals that are close to Gaussian, in addition to stabilizing the residual variances. Similar conclusions could have been reached using a goodness-of-fit test (see Section 5.2.5).

It is also possible and desirable to investigate the degree to which the residuals are uncorrelated. This question is of particular interest when the underlying data are serially correlated, which is a common condition for atmospheric variables. A simple graphical evaluation can be obtained by plotting the regression residuals as a function of time. If groups of positive and negative residuals tend to cluster together (qualitatively resembling Figure 5.4b) rather than occurring more irregularly (as in Figure 5.4a), then time correlation can be suspected.

A popular formal test for serial correlation of regression residuals, included in many computer regression packages, is the *Durbin-Watson test*. This test examines the null hypothesis that the residuals are serially independent, against the alternative that they are consistent with a first-order autoregressive process (Equation 9.16). The Durbin-Watson test statistic,

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}, \tag{7.21}$$

computes the squared differences between pairs of consecutive residuals, divided by a scaling factor proportional to the residual variance. If the residuals are positively correlated, adjacent residuals will

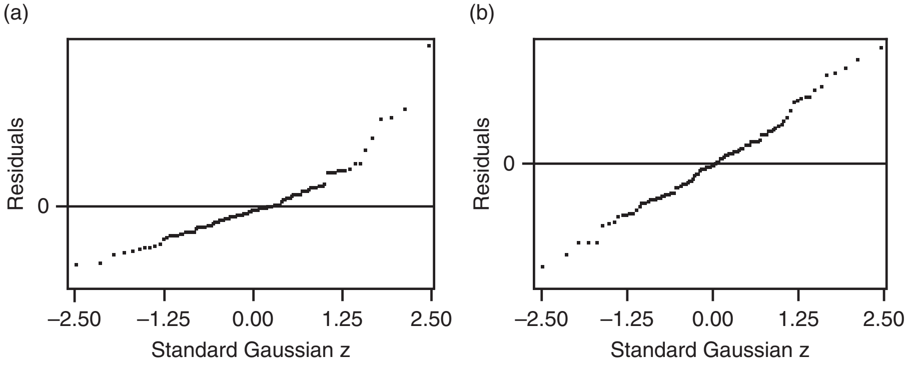


FIGURE 7.7 Gaussian quantile-quantile plots of the residuals for predictions of the untransformed predictand in Figure 7.4a (a), and the logarithmically transformed predictand in Figure 7.5b (b). In addition to producing essentially constant residual variance, logarithmic transformation of the predictand has rendered the distribution of the residuals effectively Gaussian.

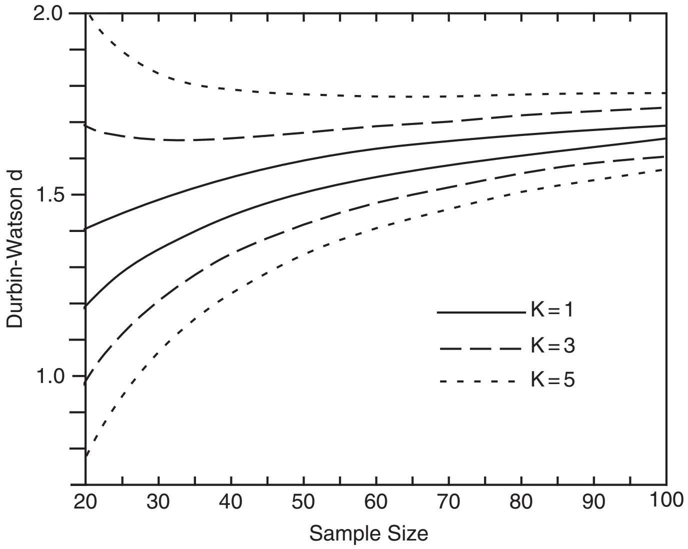


FIGURE 7.8 5%-level critical values for the Durbin-Watson statistic as a function of the sample size, for $K = 1$, 3, and 5 predictor variables. A test statistic d below the relevant lower curve results in a rejection of the null hypothesis of zero serial correlation. If the test statistic is above the relevant upper curve, the null hypothesis is not rejected. If the test statistic is between the two curves, the test is indeterminate without additional calculations.

tend to be similar in magnitude, so the Durbin-Watson statistic will be relatively small. If the sequence of residuals is randomly distributed, the sum in the numerator will tend to be larger. Therefore the null hypothesis that the residuals are independent is rejected if the Durbin-Watson statistic is sufficiently small.

Figure 7.8 shows critical values for Durbin-Watson tests at the 5% level. These vary depending on the sample size and the number of predictor (x) variables, K. For simple linear regression, $K = 1$. For each value of K, Figure 7.8 shows two curves. If the observed value of the test statistic falls below the lower curve, the null hypothesis is rejected and we conclude that the residuals exhibit significant serial correlation. If the test statistic falls above the upper curve, we do not reject the null hypothesis that the residuals are serially uncorrelated. If the test statistic falls between the two relevant curves, the test is indeterminate. The reason behind the existence of this unusual indeterminate condition is that the null distribution of the Durban-Watson statistic depends on the data set being considered. In cases where the test result is indeterminate according to Figure 7.8, some additional calculations (Durbin and Watson, 1971) can be performed to resolve the indeterminacy—that is, to find the specific location of the critical value between the appropriate pair of curves, for the particular data at hand.

Example 7.2. Examination of the Residuals from Example 7.1

A regression equation constructed using autocorrelated variables as predictand and predictor(s) does not necessarily exhibit strongly autocorrelated residuals. Consider again the regression between Ithaca and Canandaigua minimum temperatures for January 1987 in Example 7.1. The lag-1 autocorrelations (Equation 3.32) for the Ithaca and Canandaigua minimum temperature data are 0.651 and 0.672, respectively. The residuals for this regression are plotted as a function of time in Figure 7.9. A strong

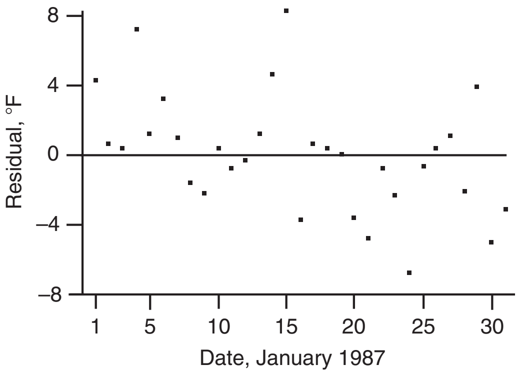


FIGURE 7.9 Residuals from the regression, Equation 7.20, plotted as a function of date. A strong serial correlation is not apparent, but the tendency for a negative slope suggests that the relationship between Ithaca and Canandaigua temperatures may be changing through the month.

serial correlation for these residuals is not apparent, and their lag-1 autocorrelation as computed using Equation 3.32 is only 0.191.

Having computed the residuals for the Canandaigua versus Ithaca minimum temperature regression, it is straightforward to compute the Durbin-Watson d (Equation 7.21). In fact, the denominator is simply the SSE from the ANOVA Table 7.2, which is 341.622. The numerator in Equation 7.21 must be computed from the residuals and is 531.36. These yield $d = 1.55$. Referring to Figure 7.8, the point at $n = 31$, $d = 1.55$ is well above the upper solid (for $K = 1$, since there is a single predictor variable) line, so the null hypothesis of uncorrelated residuals would not be rejected at the 5% level. ◇

When regression residuals are autocorrelated, statistical inferences based on their variance are degraded in the same way, and for the same reasons, that were discussed in Section 5.2.4 (Bloomfield and Nychka, 1992; Matalas and Sankarasubramanian, 2003; Santer et al., 2000; Zheng et al., 1997). In particular, positive serial correlation of the residuals leads to inflation of the variance of the sampling distribution of their sum or average, because these quantities are less consistent from batch to batch of size n. When a first-order autoregression (Equation 9.16) is a reasonable representation for these correlations (characterized by r_1), it is appropriate to apply the same variance inflation factor, $(1 + r_1)/(1 - r_1)$ (bracketed quantity in Equation 5.13), to the variance $s_e{}^2$ in, for example, Equations 7.17b and 7.18b (Matalas and Sankarasubramanian, 2003; Santer et al., 2000). The net effect is that the variance of the resulting sampling distribution is (appropriately) increased, relative to what would be calculated assuming independent regression residuals.

7.2.7. Prediction Intervals

Many times it is of interest to calculate *prediction intervals* around forecast values of the predictand (i.e., around the regression function), which are meant to bound a future value of the predictand with specified probability. When it can be assumed that the residuals follow a Gaussian distribution, it is natural to approach this problem using the unbiasedness property of the residuals (Equation 7.8), together with their estimated variance MSE $= s^2{}_e$. Using Gaussian probabilities (Table B.1), we expect a 95% prediction interval for a future residual, or specific future forecast, to be approximately bounded by $\hat{y} \pm 2s_e$.

The $\pm 2s_e$ rule of thumb is often a quite good approximation to the width of a true 95% prediction interval, especially when the sample size is large. However, because both the sample mean of the predictand and the slope of the regression are subject to sampling variations, the prediction variance for future data (i.e., for data not used in the fitting of the regression) is somewhat larger than the regression MSE. For a forecast of y using the predictor value x_0, this prediction variance is given by

$$s_{\hat{y}}^2 = s_e^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]. \tag{7.22}$$

That is, the prediction variance is proportional to the regression MSE, but is larger to the extent that the second and third terms inside the square brackets are appreciably larger than zero. The second term derives from the uncertainty in estimating the true mean of the predictand from a finite sample of size n (compare Equation 5.4), and becomes much smaller than one for large sample sizes. The third term derives from the uncertainty in estimation of the slope (it is similar in form to Equation 7.18b), and indicates that predictions far removed from the center of the data used to fit the regression will be more uncertain than predictions made near the sample mean. However, even if the numerator in this third term is fairly large, the term itself will tend to be small if a large data sample was used to construct the regression equation, since there are n non-negative terms of generally comparable magnitude in the denominator.

It is sometimes also of interest to compute *confidence intervals* for the regression function itself. These will be narrower than the prediction intervals for future individual data values, reflecting a smaller variance in a way that is analogous to the variance of a sample mean being smaller than the variance of the underlying data values. The variance for the sampling distribution of the regression function, or equivalently the variance of the conditional mean of the predictand given a particular predictor value x_0, is

$$s_{\bar{y}|x_0}^2 = s_e^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]. \tag{7.23}$$

This expression is similar to Equation 7.22 but is smaller by the amount $s^2{}_e$. That is, there are contributions to this variance due to uncertainty in the mean of the predictand (or, equivalently the vertical position of the regression line, or the intercept), corresponding to the first of the two terms in the square brackets; and to uncertainty in the slope, corresponding to the second term. There is no contribution to Equation 7.23 reflecting scatter of data around the regression line, which is the difference between Equations 7.22 and 7.23. The extension of Equation 7.23 for multiple regression is given in Equation 10.41.

Figure 7.10 compares prediction and confidence intervals computed using Equations 7.22 and 7.23, in the context of the regression from Example 7.1. Here the regression (Equation 7.20) fit to the 31 data points (dots) is shown by the heavy solid line. The 95% prediction interval around the regression computed as $\pm 1.96\ s_{\hat{y}}$, using the square root of Equation 7.22, is indicated by the pair of slightly curved solid black lines. As noted earlier, these bounds are only slightly wider than those given by the simpler approximation $\hat{y} \pm 1.96\ s_e$ (dashed lines) because the second and third terms in the square brackets of Equation 7.22 are relatively small, even for moderate n. The pair of gray curved lines locate the 95% confidence interval for the conditional mean of the predictand. These are much narrower than the prediction interval because they account only for sampling variations in the regression parameters, without direct contributions from the prediction variance $s^2{}_e$.

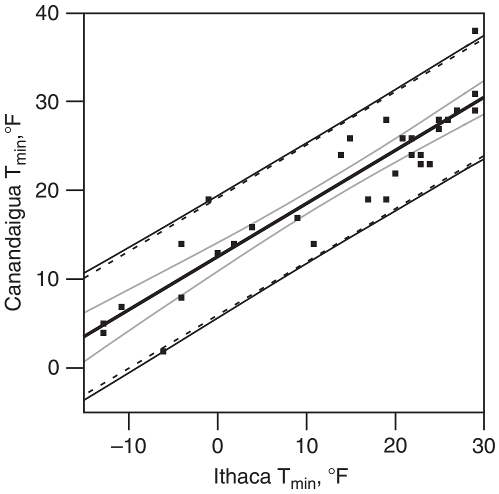


FIGURE 7.10 Prediction and confidence intervals around the regression derived in Example 7.1 (thick black line). Light solid lines indicate 95% prediction intervals for future data, computed using Equation 7.22, and the corresponding dashed lines simply locate the predictions ± 1.96 s_e. Light gray lines locate 95% confidence intervals for the regression function (Equation 7.23). Data to which the regression was fit are also shown.

Equations 7.17 through 7.19 define the parameters of a bivariate normal distribution for the two regression parameters. Imagine using the methods outlined in Section 4.7 to generate pairs of intercepts and slopes according to that distribution, and therefore to generate realizations of plausible regression lines. One interpretation of the gray curves in Figure 7.10 is that they would contain 95% of those regression lines (or, equivalently, 95% of the regression lines computed from different samples of data of this kind, each with size $n = 31$). The minimum separation between the gray curves (at the average Ithaca T_{min} = 13°F) reflects the uncertainty in the intercept. Their spreading at more extreme temperatures reflects the fact that uncertainty in the slope (i.e., uncertainty in the angle of the regression line) will produce more uncertainty in the conditional expected value of the predictand at the extremes than near the mean, because any regression line must pass through the point located by the two sample means.

The result of Example 7.2 is that the residuals for this regression can reasonably be regarded as independent. Also, some of the sample lag-1 autocorrelation of $r_1 = 0.191$ can be attributed to the time trend evident in Figure 7.9. However, if the residuals are significantly correlated, and the nature of that correlation is plausibly represented by a first-order autoregression (Equation 9.16), it would be appropriate to increase the residual variances s^2_e in Equations 7.22 and 7.23 by multiplying them by the variance inflation factor $(1 + r_1)/(1 - r_1)$.

Special care is required when computing prediction and confidence intervals for regressions involving transformed predictands. For example, if the relationship shown in Figure 7.5a (involving a log-transformed predictand) were to be used in forecasting, dimensional values of the predictand would need to be recovered in order to make the forecasts interpretable. That is, the predictand ln ($\hat{y}$) would need to be back-transformed, yielding the forecast $\hat{y} = \exp[\ln(\hat{y})] = \exp[a + bx]$. Similarly, the limits of the prediction intervals would also need to be back-transformed. For example the 95% prediction interval would be approximately $\ln(\hat{y}) \pm 1.96\ s_e$, because the regression residuals and their assumed Gaussian distribution pertain to the transformed predictand values. The lower and upper limits of this interval, when expressed on the original untransformed scale of the predictand, would be approximately $\exp[a + bx - 1.96\ s_e]$ and $\exp[a + bx + 1.96 s_e]$. These limits would not be symmetrical around $\hat{y}$ and would extend further for the larger values, consistent with the longer right tail of the predictand distribution.

Equations 7.22 and 7.23 are valid for simple linear regression. The corresponding equations for multiple regression are similar, but are more conveniently expressed in matrix algebra notation (e.g., Draper and Smith, 1998; Neter et al., 1996). As is the case for simple linear regression, the prediction variance is quite close to the MSE for moderately large samples.

7.2.8. Multiple Linear Regression

Multiple linear regression is the more general (and more common) situation of linear regression. As in the case of simple linear regression, there is still a single predictand, y, but in distinction there is more than one predictor (x) variable. The preceding treatment of simple linear regression was relatively lengthy, in part because most of what was presented generalizes readily to the case of multiple linear regression.

Let K denote the number of predictor variables. Simple linear regression is then the special case of $K = 1$. The prediction equation (corresponding to Equation 7.1) becomes

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_Kx_K. \tag{7.24}$$

Each of the K predictor variables has its own coefficient, analogous to the slope, b, in Equation 7.1. For notational convenience, the intercept (or *regression constant*) is denoted as b_0 rather than a, as in Equation 7.1. These $K + 1$ regression coefficients often are called the *regression parameters*.

Equation 7.2 for the residuals is still valid, if it is understood that the predicted value $\hat{y}$ is a function of a vector of predictors, x_k, $k = 1, \ldots, K$. If there are $K = 2$ predictor variables, the residual can still be visualized as a vertical distance. In that case the regression function (Equation 7.24) is a surface rather than a line, and the residual corresponds geometrically to the distance above or below this surface along a line perpendicular to the (x_1, x_2) plane. The geometric situation is analogous for $K \geq 3$, but is not easily visualized. Also in common with simple linear regression, the average residual is guaranteed to be zero, so that the residual distributions are centered on the predicted values $\hat{y}_i$. Accordingly, these predicted values can be regarded as conditional means given particular values for a set of K predictors.

The $K + 1$ parameters in Equation 7.24 are found, as before, by minimizing the sum of squared residuals. This is achieved by simultaneously solving $K + 1$ equations analogous to Equation 7.5. This minimization is most conveniently done using matrix algebra, the details of which can be found in standard regression texts (e.g., Draper and Smith, 1998; Neter et al., 1996). The basics of the process are outlined in Example 10.2. In practice, the calculations usually are done using statistical software. They are again summarized in an ANOVA table, of the form shown in Table 7.3. As before, SST is computed using Equation 7.12, SSR is computed using Equation 7.13a, and SSE is computed using the difference SST – SSR. The sample variance of the residuals is MSE = SSE/(n – K – 1). The coefficient of determination is computed according to Equation 7.16, although it is no longer the square of the Pearson correlation coefficient between the predictand and any of the predictor variables. The procedures presented previously for examination of residuals are applicable to multiple regression as well.

7.2.9. Derived Predictor Variables in Multiple Regression

Multiple regression opens up the possibility of an essentially unlimited number of potential predictor variables. An initial list of potential predictor variables can be expanded manyfold by also considering nonlinear mathematical transformations of these variables as potential predictors. The derived

TABLE 7.3 Generic analysis of variance (ANOVA) table for multiple linear regression. Table 7.1 for simple linear regression can be viewed as a special case, with $K = 1$.

Source	df	SS	MS	F
Total	$n - 1$	SST		
Regression	K	SSR	MSR = SSR / K	F = MSR/MSE
Residual	$n - K - 1$	SSE	MSE = SSE/($n - K - 1$)=s_e^2	

predictors must be nonlinear functions of the primary predictors in order for the computations (in particular, for the matrix inversion indicated in Equation 10.39) to be possible. Such *derived predictors* can be very useful in producing a good regression equation.

In some instances the most appropriate forms for predictor transformations may be suggested by physical understanding of the data-generating process. In the absence of a strong physical rationale for particular predictor transformations, the choice of a transformation or set of transformations may be made purely empirically, perhaps by subjectively evaluating the general shape of the point cloud in a scatterplot, or the nature of the deviation of a residual plot from its ideal form. For example, the curvature in the residual plot in Figure 7.6d suggests that addition of the derived predictor $x_2 = x_1{}^2$ might improve the regression relationship. It may happen that the empirical choice of a transformation for a predictor variable in regression leads to a greater physical understanding, which is a highly desirable outcome in a research setting. This outcome would be less important in a purely forecasting setting, where the emphasis is on producing good forecasts rather than knowing precisely why the forecasts are good.

Transformations such as $x_2 = x_1{}^2$, $x_2 = \sqrt{x_1}$, $x_2 = 1/x_1$, or any other power transformation of an available predictor, can be regarded as another potential predictor. Similarly, trigonometric (sine, cosine, etc.), exponential or logarithmic functions, or combinations of these are useful in some situations. Another commonly used transformation is to a *binary variable*, or *dummy variable*. Binary variables take on one of two values (usually 0 and 1, although the particular choices do not affect subsequent use of the regression equation), depending on whether the variable being transformed is above or below a threshold or cutoff, c. That is, a binary variable x_2 could be constructed from another predictor x_1 according to the transformation

$$x_2 = \begin{cases} 1, & if\ x_1 > c \\ 0, & if\ x_1 \le c \end{cases}. \tag{7.25}$$

More than one binary predictor can be constructed from a single x_1 by choosing different values of the cutoff, c, for x_2, x_3, x_4, and so on.

Even though transformed variables may be nonlinear functions of other variables, the overall framework is still known as multiple linear regression. Once a derived variable has been defined it is just another variable, regardless of how the transformation was made. More formally, the “linear” in multiple linear regression refers to the regression equation being linear in the parameters, b_k.

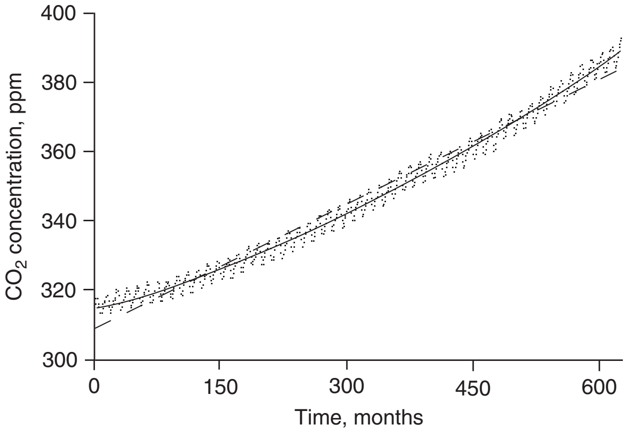


FIGURE 7.11 The Keeling Mauna Loa monthly CO_2 concentration data (March 1958–May 2010), with linear (dashed) and quadratic (solid) least-squares fits.

Example 7.3. A Multiple Regression with Derived Predictor Variables

Figure 7.11 shows a scatterplot of the famous Keeling monthly-averaged carbon dioxide (CO_2) concentration data from Mauna Loa in Hawaii for the period March 1958 through May 2010. Representing the obvious time trend as a straight line yields the regression results shown in Table 7.4a, and the regression line is also plotted (dashed) in Figure 7.11. The results indicate a strong time trend, with the calculated standard error for the slope being much smaller than the estimated slope. The intercept merely estimates the CO_2 concentration at $t = 0$, or February 1958, so the implied test for its difference from zero is of no interest. A literal interpretation of the MSE would suggest that a 95% prediction interval for measured CO_2 concentrations around the regression line would be about $\pm 2\sqrt{\text{MSE}} = \pm 6.6$ ppm.

However, examination of a plot of the residuals versus time for this linear regression would reveal a bowing pattern similar to that in Figure 7.6d, with a tendency for positive residuals at the beginning and end of the record and with negative residuals being more common in the central part of the record. This can be discerned from Figure 7.11 by noticing that most of the points fall above the dashed line early and late in the record, and fall below the line toward the middle.

This problem with the residuals can be alleviated (and the regression consequently improved) by fitting a quadratic curve to the time trend. To do this, a second predictor is added to the regression, and that predictor is simply the square of the time variable. That is, a multiple regression with $K = 2$ is fit using the predictors $x_1 = t$ and $x_2 = t^2$. Once defined, x_2 is just another predictor variable, taking on values between 1^2 and $627^2 = 393{,}129$. The resulting least-squares quadratic regression is shown by the solid curve in Figure 7.11, and the corresponding regression statistics are summarized in Table 7.4b.

Of course the SST in Tables 7.4a and 7.4b are the same since both pertain to the same predictand, the CO_2 concentrations. For the quadratic regression, both the coefficients $b_1 = 0.0663$ and $b_2 = 0.00008528$ are substantially larger than their respective standard errors. The value of $b_0 = 314.3$ is again just the estimate of the CO_2 concentration at $t = 0$, and judging from the scatterplot this intercept is a better estimate of its true value than was obtained from the simple linear regression. The data points are fairly evenly scattered around the quadratic trend line throughout the time period, so the

TABLE 7.4 ANOVA tables and regression summaries for three regressions fit to the 1958–2010 Keeling CO_2 data in Figure 7.11. The variable *t* (time) is a consecutive numbering of the months, with March 1958 = 1 and May 2010 = 627. There are *n* = 620 data points and 7 missing months.

(a) Linear Fit

Source	df	SS	MS	F
Total	619	297772		
Regression	1	290985	290985	26497
Residual	618	6786.8	10.982	

Variable	Coefficient	s.e.	t-ratio
Constant	308.6	0.2687	1148
t	0.1201	0.0007	163.0

(b) Quadratic Fit

Source	df	SS	MS	F
Total	619	297772		
Regression	2	294817	147409	30781
Residual	617	2954.8	4.789	

Variable	Coefficient	s.e.	t-ratio
Constant	314.3	0.2687	1170
t	0.0663	0.0020	33.8
t^2	.00008528	0.0000	28.3

(c) Including quadratic trend, and harmonic terms to represent the annual cycle

Source	df	SS	MS	F
Total	619	297772		
Regression	4	297260	74314.9	89210
Residual	615	512.31	0.83303	

Variable	Coefficient	s.e.	t-ratio
Constant	314.2	0.1121	2804
t	0.0669	0.0008	81.7
t^2	0.00008439	0.0000	67.1
$\cos(2\pi t/12)$	1.122	0.0518	21.6
$\sin(2\pi t/12)$	2.573	0.0518	49.6

residual plot would exhibit the desired horizontal banding. Using this analysis, an approximate 95% prediction interval of $\pm 2\sqrt{\text{MSE}} = \pm 4.4$ ppm for CO_2 concentrations around the quadratic regression would be inferred throughout the range of these data.

The quadratic function of time provides a reasonable approximation of the annual-average CO_2 concentration for the 53 years represented by the regression, although we can find periods of time where the center of the point cloud wanders away from the curve. More importantly, however, a close inspection of the data points in Figure 7.11 reveals that they are not scattered randomly around the quadratic time trend. Rather, they execute a regular, nearly sinusoidal variation around the quadratic curve that is evidently an annual cycle. The resulting serial correlation in the residuals can easily be detected using the Durbin-Watson statistic, $d = 0.135$ (compare Figure 7.8). The CO_2 concentrations are lower in late summer and higher in late winter as a consequence of the annual cycle of photosynthetic carbon uptake by northern hemisphere land plants and carbon release from the decomposing dead plant parts. As will be shown in Section 9.4.2, this regular 12-month variation can be represented by introducing two more derived predictor variables into the equation, $x_3 = \cos(2\pi t\ /12)$ and $x_4 = \sin(2\pi t\ /12)$. Notice that both of these derived variables are functions only of the time variable t.

Table 7.4c indicates that, together with the linear and quadratic predictors included previously, these two harmonic predictors produce a very close fit to the data. The resulting prediction equation is

$$[CO_2] = \underset{(.1121)}{314.2} + \underset{(.0008)}{0.0669}\, t + \underset{(.0000)}{.00008438}\ t^2 + \underset{(.0518)}{1.122}\ \cos\left(\frac{2\pi t}{12}\right) + \underset{(.0518)}{2.573}\ \sin\left(\frac{2\pi t}{12}\right), \tag{7.26}$$

with all regression coefficients being much larger than their respective standard errors. The near equality of SST and SSR indicate that the predicted values are nearly coincident with the observed CO_2 concentrations (compare Equations 7.12 and 7.13a). The resulting coefficient of determination is $R^2 = 297260/297772 = 99.83\%$, and the approximate 95% prediction interval implied by $\pm 2\ \sqrt{\text{MSE}}$ is only ± 1.8 ppm. A graph of Equation 7.26 would wiggle up and down around the solid curve in Figure 7.11, passing rather close to each of the data points. ◇

7.3. NONLINEAR REGRESSION

7.3.1. Generalized Linear Models

Although linear least-squares regression accounts for the overwhelming majority of regression applications, it is also possible to fit regression functions that are nonlinear (in the regression parameters). Nonlinear regression can be appropriate when a nonlinear relationship is dictated by the nature of the physical problem at hand, and/or the usual assumptions of Gaussian residuals with constant variance are untenable. In these cases the fitting procedure is usually iterative and based on maximum-likelihood methods (see Section 4.6).

This section introduces two such regression structures, both of which are important examples of a class of nonlinear statistical models known as *generalized linear models* (GLMs) (McCullagh and Nelder, 1989). Generalized linear models extend linear statistical models, such as multiple linear regression, by representing the predictand as a nonlinear function of a linear regression function. The nonlinearity is represented by a 1-to-1 (and therefore invertible) function known as the *link function*, $g(\hat{y})$. Accordingly, the GLM extension of the ordinary linear multiple regression (Equation 7.24) is

$$g(\hat{y}) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_Kx_K, \tag{7.27}$$

where the specific form of the link function is chosen according to the nature of the predictand data. Comparing Equation 7.27 and 7.24 shows that ordinary linear regression is a special case of a GLM, with the identity link, that is, $g(\hat{y}) = \hat{y}$. Because the link function will be invertible, GLM equations are often written equivalently as

$$\hat{y} = g^{-1}(b_0 + b_1x_1 + b_2x_2 + \cdots + b_Kx_K). \tag{7.28}$$

7.3.2. Logistic Regression

One important advantage of statistical over (deterministic) dynamical forecasting methods is the capacity to produce probability forecasts. Inclusion of probability elements into the forecast format is advantageous because it provides an explicit expression of the inherent uncertainty or state of knowledge about the future weather, and because probabilistic forecasts allow users to extract more value from them when making decisions (e.g., Katz and Murphy, 1997a,b; Krzysztofowicz, 1983; Murphy, 1977; Thompson, 1962). In a sense, ordinary linear regression produces probability information about a predictand, for example, through the 95% prediction interval around the regression function given by the $\pm 2\sqrt{\text{MSE}}$ rule. More narrowly, however, probability forecasts are forecasts for which the predictand is a probability, rather than the value of a physical variable.

Most commonly, systems for producing probability forecasts are developed in a regression setting by first transforming the predictand to a binary (or dummy) variable, taking on the values zero and one. That is, regression procedures are implemented after applying Equation 7.25 to the predictand, y, rather than to a predictor. In a sense, zero and one can be viewed as probabilities of the dichotomous event not occurring or occurring, respectively, after it has been observed.

The simplest approach to regression when the predictand is binary is to use the machinery of ordinary multiple regression as described in the previous section. In the meteorological literature this is called Regression Estimation of Event Probabilities (REEP) (Glahn, 1985). The main justification for the use of REEP is that it is no more computationally demanding than the fitting of any other linear regression, and so it has been used extensively when computational resources have been limiting. The resulting predicted values are usually between zero and one, and it has been found through operational experience that these predicted values can usually be treated as specifications of probabilities for the event $\{Y = 1\}$. However, one obvious problem with REEP is that some of the resulting forecasts may not lie on the unit interval, particularly when the predictands are near the limits, or outside, of their ranges in the training data. This logical inconsistency usually causes little difficulty in an operational setting because multiple-regression forecast equations with many predictors rarely produce such nonsense probability estimates. When the problem does occur, the forecast probability is usually near zero or one, and the operational forecast can be issued as such.

Two other difficulties associated with forcing a linear regression onto a problem with a binary predictand are that the residuals are clearly not Gaussian and their variances are not constant. Because the predictand can take on only one of two values, a given regression residual can also take on only one of two values, and so the residual distributions are Bernoulli (i.e., binomial, Equation 4.1, with $N = 1$). Furthermore, the variance of the residuals is not constant, but depends on the ith predicted probability p_i according to $(p_i)(1 - p_i)$. It is possible to simultaneously bound the regression estimates for binary predictands on the interval (0, 1) and to accommodate the Bernoulli distributions for the regression residuals, using a technique known as *logistic regression*. Some recent examples of logistic regression in the atmospheric science literature are Applequist et al. (2002), Buishand et al. (2004), Hilliker and Fritsch (1999), Lehmiller et al. (1997), Mazany et al. (2002), Watson and Colucci (2002), and Wilks (2009).

Logistic regressions are fit to binary predictands using the log-odds, or *logit*, link function $g(p) = \ln[p/(1 - p)]$, yielding the generalized linear model

$$\ln\left(\frac{p_i}{1 - p_i}\right) = b_0 + b_1 x_1 + \cdots + b_K x_K, \tag{7.29a}$$

which can also be expressed in the form of Equation 7.28 as

$$p_i = \frac{\exp(b_0 + b_1 x_1 + \cdots + b_K x_K)}{1 + \exp(b_0 + b_1 x_1 + \cdots + b_K x_K)} = \frac{1}{1 + \exp(-b_0 - b_1 x_1 - \cdots - b_K x_K)}. \tag{7.29b}$$

Here the predicted value p_i results from the ith set of predictors (x_1, x_2, . . ., x_K) of n such sets. Geometrically, logistic regression is most easily visualized for the single-predictor case ($K = 1$), for which Equation 7.29b is an S-shaped curve that is a function of x_1. In the limits, $b_0 + b_1 x_1 \to +\infty$ results in the exponential function in the first equality of Equation 7.29b becoming arbitrarily large so that the predicted value p_i approaches one. As $b_0 + b_1 x_1 \to -\infty$, the exponential function approaches zero, and thus so does the predicted value. Depending on the parameters b_0 and b_1, the function rises gradually or abruptly from zero to one (or falls, for $b_1 < 0$, from one to zero) at intermediate values of x_1. Thus it is guaranteed that logistic regression will produce properly bounded probability estimates. The logistic function is convenient mathematically, but it is not the only function that could be used in this context. Another alternative yielding a very similar shape involves using the inverse Gaussian CDF for the link function, yielding $p_i = \Phi(b_0 + b_1 x_1 + ...+ b_K x_K)$, which is known as *probit regression*.

Equation 7.29a shows that logistic regression can be viewed as linear in terms of the logarithm of the odds ratio, $p_i / (1 - p_i)$. Superficially it appears that Equation 7.29a could be fit using ordinary linear regression, except that the predictand is binary, so the left-hand side will be either ln(0) or ln(∞). However, fitting the regression parameters can be accomplished using the method of maximum likelihood, recognizing that the residuals are Bernoulli variables. Assuming that Equation 7.29 is a reasonable model for the smooth changes in the probability of the binary outcome as a function of the predictors, the probability distribution function for the ith residual is Equation 4.1, with $N = 1$ and p_i as specified by Equation 7.29b. The corresponding likelihood is of the same functional form, except that the values of the predictand y and the predictors x are fixed, and the probability p_i is the variable. If the ith residual corresponds to a success (i.e., the event occurs, so $y_i = 1$), the likelihood is $\Lambda = p_i$ (as specified in Equation 7.29b), and otherwise $\Lambda = 1 - p_i = 1/(1 + \exp[b_0 + b_1 x_1 +. ..+ b_K x_K]\ )$. If the n sets of observations (predictand and predictor(s)) are independent, the joint likelihood for the $K + 1$ regression parameters is simply the product of the n individual likelihoods, or

$$\Lambda(\mathbf{b}) = \prod_{i=1}^{n} \frac{y_i \exp(b_0 + b_1 x_1 + \cdots + b_K x_K) + (1 - y_i)}{1 + \exp(b_0 + b_1 x_1 + \cdots + b_K x_K)}. \tag{7.30}$$

Since the y's are binary [0, 1] variables, each factor in Equation 7.30 for which $y_i = 1$ is equal to p_i (Equation 7.29b), and the factors for which $y_i = 0$ are equal to $1 - p_i$. As usual, it is more convenient to estimate the regression parameters by maximizing the log-likelihood

$$L(\mathbf{b}) = \ln[\Lambda(\mathbf{b})] = \sum_{i=1}^{n} \{y_i(b_0 + b_1 x_1 + \cdots b_K x_K) - \ln[1 + \exp(b_0 + b_1 x_1 + \cdots b_K x_K)]\} \tag{7.31}$$

The combinatorial factor in Equation 4.1 has been omitted because it does not involve the unknown regression parameters, and so will not influence the process of locating the maximum of the function. Usually statistical software will be used to find the values of the b's maximizing this function, using iterative methods such as those in Section 4.6.2 or 4.6.3.

Some software will display information relevant to the strength of the maximum-likelihood fit using what is called the *analysis of deviance* table, which is analogous to the ANOVA table (see Table 7.3) for linear regression. More about analysis of deviance can be learned from sources such as Healy (1988) or McCullagh and Nelder (1989), although the idea underlying an analysis of deviance table is the likelihood ratio test (Equation 5.19). As more predictors and thus more regression parameters are added to Equation 7.29, the log-likelihood will progressively increase as more latitude is provided to accommodate the data. Whether that increase is sufficiently large to reject the null hypothesis that a particular, smaller, regression equation is adequate, is judged in terms of twice the difference of the log-likelihoods relative to the χ^2 distribution, with degrees of freedom v equal to the difference in numbers of parameters between the null-hypothesis regression and the more elaborate regression being considered.

The likelihood ratio test is appropriate when a single candidate logistic regression is being compared to a null model. Often H_0 will specify that all the regression parameters except b_0 are zero, in which case the question being addressed is whether the predictors $\boldsymbol{x}$ being considered are justified in favor of the constant (no-predictor) model with $b_0 = \ln\,[\Sigma y_i\,/n\,/\,(1 - \Sigma y_i/n)\,]$. However, if multiple alternative logistic regressions are being entertained, computing the likelihood ratio test for each alternative raises the problem of test multiplicity (see Section 5.4.1). In such cases it is better to compute either the Bayesian Information Criterion (BIC) statistic (Schwarz, 1978)

$$BIC = -2\,L(\mathbf{b}) + (K+1)\ln(n) \tag{7.32}$$

or the Akaike Information Criterion (AIC) (Akaike, 1974)

$$AIC = -2\,L(\mathbf{b}) + 2(K+1) \tag{7.33}$$

for each candidate model. Both the AIC and BIC statistics consist of twice the negative of the log-likelihood plus a penalty for the number of parameters fit, and the preferred regression will be the one minimizing the chosen criterion. The BIC statistic will generally be better for large-n problems since its probability of selecting the proper member of the class of models considered approaches 1 as n—>∞; whereas for smaller sample sizes BIC often chooses models that are simpler than are justified by the data, in which cases AIC may be preferred.

Example 7.4. Comparison of REEP and Logistic Regression

Figure 7.12 compares the results of REEP (dashed) and logistic regression (solid) for some of the January 1987 data from Table A.1. The predictand is daily Ithaca precipitation, transformed to a binary variable using Equation 7.25 with $c = 0$. That is, $y = 0$ if the precipitation is zero, and $y = 1$ otherwise. The predictor is the Ithaca minimum temperature for the same day. The REEP (linear regression) equation has been fit using ordinary least squares, yielding $b_0 = 0.208$ and $b_1 = 0.0212$. This equation specifies negative probability of precipitation if the temperature predictor is less than about –9.8°F and specifies probability of precipitation greater than one if the minimum temperature is greater than about 37.4°F. The parameters for the logistic regression, fit using maximum likelihood,

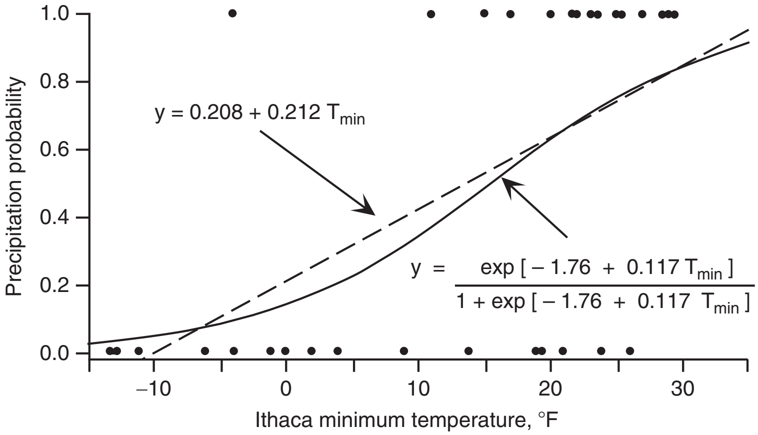


FIGURE 7.12 Comparison of regression probability forecasting using REEP (dashed) and logistic regression (solid) using the January 1987 data set in Table A.1. The linear function was fit using least squares, and the logistic curve was fit using maximum likelihood, to the data shown by the dots. The binary predictand $y = 1$ if Ithaca precipitation is greater than zero, and $y = 0$ otherwise.

are $b_0 = –1.76$ and $b_1 = 0.117$. The logistic regression curve produces probabilities that are similar to the REEP specifications through most of the temperature range, but are constrained by the functional form of Equation 7.29 to lie between zero and one, even for extreme values of the predictor.

Maximizing Equation 7.31 for logistic regression with a single ($K = 1$) predictor is simple enough that the Newton-Raphson method (see Section 4.6.2) can be implemented easily and is reasonably robust to poor initial guesses for the parameters. The counterpart to Equation 4.76 for this problem is

$$\begin{bmatrix} b_0^* \\ b_1^* \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} - \begin{bmatrix} \sum_{i=1}^{n} (p_i^2 - p_i) & \sum_{i=1}^{n} x_i (p_i^2 - p_i) \\ \sum_{i=1}^{n} x_i (p_i^2 - p_i) & \sum_{i=1}^{n} x_i^2 (p_i^2 - p_i) \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} (y_i - p_i) \\ \sum_{i=1}^{n} x_i (y_i - p_i) \end{bmatrix}, \qquad (7.34)$$

where p_i is a function of the regression parameters b_0 and b_1, and depends also on the predictor data x_i, as shown in Equation 7.29b. The first derivatives of the log-likelihood (Equation 7.31) with respect to b_0 and b_1 are in the vector enclosed by the rightmost square brackets, and the second derivatives are contained in the matrix to be inverted. Beginning with an initial guess for the parameters (b_0, b_1), updated parameters (b^*_0, b^*_1) are computed and then resubstituted into the right-hand side of Equation 7.34 for the next iteration. For example, assuming initially that the Ithaca minimum temperature is unrelated to the binary precipitation outcome, so $b_0 = –0.0645$ (the log of the observed odds ratio, for constant $p = 15/31$) and $b_1 = 0$; the updated parameters for the first iteration are $b^*_0 = –0.0645 – (–0.251)(–0.000297) – (0.00936)(118.0) = –1.17$, and $b^*_1 = 0 – (0.00936)(–0.000297) – (–0.000720)(118.0) = 0.085$. These updated parameters increase the log-likelihood from –21.47 for the constant model (calculated using Equation 7.31, imposing $b_0 = –0.0645$ and $b_1 = 0$), to –16.00. After four iterations the algorithm has converged, with a final (maximized) log-likelihood of –15.67.

Is the logistic relationship between Ithaca minimum temperature and the probability of precipitation statistically significant? This question can be addressed using the likelihood ratio test (Equation 5.19). The appropriate null hypothesis is that $b_1 = 0$, so $L(H_0) = –21.47$, and $L(H_A) = –15.67$ for the fitted regression. If H_0 is true, then the observed test statistic $\Lambda^* = 2\,[L(H_A) - L(H_0)\,] = 11.6$ is a realization from the χ^2 distribution with $\nu = 1$ (the difference in the number of parameters between the two regressions), and the test is one-tailed because small values of the test statistic are favorable to H_0. Referring to the first row of Table B.3, it is clear that the regression is significant at the 0.1% level.◇

7.3.3. Poisson Regression

Another regression setting where the residual distribution may be poorly represented by the Gaussian is the case where the predictand consists of counts; that is, each of the y's is a non-negative integer. Particularly if these counts tend to be small, the residual distribution is likely to be asymmetric, and we would like a regression predicting these data to be incapable of implying nonzero probability for negative counts.

A natural probability model for count data is the Poisson distribution (Equation 4.11). Recall that one interpretation of a regression function is as the conditional mean of the predictand, given specific value(s) of the predictor(s). If the outcomes to be predicted by a regression are Poisson-distributed counts, but the Poisson parameter μ may depend on one or more predictor variables, we can structure a regression to specify the Poisson mean as a nonlinear function of those predictors using the link function $g(\mu) = \ln(\mu)$. The resulting GLM can then be written as

$$\ln(\mu_i) = b_0 + b_1 x_1 + \cdots + b_K x_K, \tag{7.35a}$$

or

$$\mu_i = \exp[b_0 + b_1 x_1 + \cdots + b_K x_K]. \tag{7.35b}$$

Equation 7.35 is not the only function that could be used for this purpose, but framing the problem in this way makes the subsequent mathematics quite tractable, and the logarithmic link function ensures that the predicted Poisson mean is non-negative. Some applications of Poisson regression are described in Elsner and Schmertmann (1993), Elsner et al. (2001), McDonnell and Holbrook (2004), Paciorek et al. (2002), Parisi and Lund (2008), and Solow and Moore (2000).

Having framed the regression in terms of Poisson distributions for the y_i conditional on the corresponding set of predictor variables $\boldsymbol{x}_i = \{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_K\}$, the natural approach to parameter fitting is to maximize the Poisson log-likelihood, written in terms of the regression parameters. Again assuming independence among the n data values, the log-likelihood is

$$L(\mathbf{b}) = \sum_{i=1}^{n} \{ y_i (b_0 + b_1 x_1 + \cdots + b_K x_K) - \exp(b_0 + b_1 x_1 + \cdots + b_K x_K) \}, \tag{7.36}$$

where the term involving $y!$ from the denominator of Equation 4.11 has been omitted because it does not involve the unknown regression parameters, and so will not influence the process of locating the maximum of the function. An analytic maximization of Equation 7.36 in general is not possible, so that statistical software will approximate the maximum iteratively, typically using one of the methods outlined in Section 4.6.2 or 4.6.3. For example, if there is a single ($K = 1$) predictor, the Newton-Raphson method (see Section 4.6.2) iterates the solution according to

$$\begin{bmatrix} b_0^* \\ b_1^* \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} - \begin{bmatrix} -\sum_{i=1}^{n} \mu_i & -\sum_{i=1}^{n} x_i \mu_i \\ -\sum_{i=1}^{n} x_i \mu_i & -\sum_{i=1}^{n} x_i^2 \mu_i \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} (y_i - \mu_i) \\ \sum_{i=1}^{n} x_i (y_i - \mu_i) \end{bmatrix}, \tag{7.37}$$

where μ_i is the conditional mean as a function of the ith set of regression parameters as defined in Equation 7.35b. Equation 7.37 is the counterpart of Equation 4.76 for fitting the gamma distribution, and Equation 7.34 for logistic regression.

Example 7.5. A Poisson Regression

Consider the annual counts of tornados reported in New York State for 1959–1988 in Table 7.5. Figure 7.13 shows a scatterplot of these as a function of average July temperatures at Ithaca in the corresponding years. The solid curve is a Poisson regression function, and the dashed line shows the ordinary linear least-squares linear fit. The nonlinearity of the Poisson regression is quite modest over the range of the training data, although the regression function would remain strictly positive regardless of the magnitude of the predictor variable.

The relationship is weak but slightly negative. The significance of the Poisson regression usually would be judged using the likelihood ratio test (Equation 5.19). The maximized log-likelihood (Equation 7.36) is 74.26 for $K = 1$, whereas the log-likelihood with only the intercept $b_0 = \ln(\Sigma y/n) = 1.526$ is 72.60. Comparing $\Lambda^* = 2$ (74.26 – 72.60) = 3.32 to the χ^2 distribution quantiles in Table B.3 with $v = 1$ (the difference in the number of fitted parameters) indicates that b_1 would be judged significantly different from zero at the 10% level, but not at the 5% level. For the linear regression, the t ratio for the slope parameter b_1 is –1.86, implying a two-tailed p value of 0.068, which is an essentially equivalent result.

The primary difference between the Poisson and linear regressions in Figure 7.13 is in the residual distributions, and therefore in the probability statements about the specified predicted values.

TABLE 7.5 Numbers of tornados reported annually in New York State, 1959–1988.

1959	3	1969	7	1979	3
1960	4	1970	4	1980	4
1961	5	1971	5	1981	3
1962	1	1972	6	1982	3
1963	3	1973	6	1983	8
1964	1	1974	6	1984	6
1965	5	1975	3	1985	7
1966	1	1976	7	1986	9
1967	2	1977	5	1987	6
1968	2	1978	8	1988	5

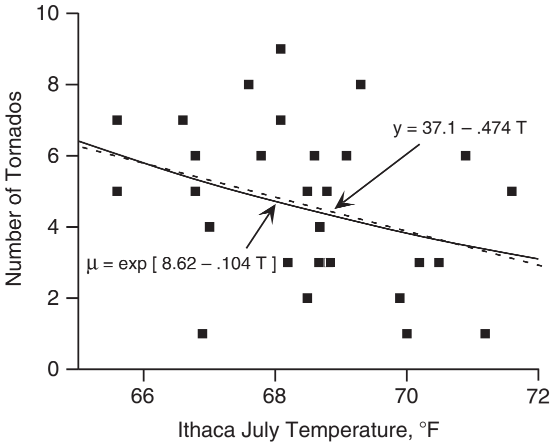


FIGURE 7.13 Annual New York tornado counts, 1959–1988 (Table 7.5), as a function of average Ithaca July temperatures in the same year. The solid curve shows the Poisson regression fit using maximum likelihood (Equation 7.37), and the dashed line shows the ordinary least-squares linear regression.

Consider, for example, the number of tornados specified when $T = 70°F$. For the linear regression, $\hat{y} = 3.92$ tornados, with a Gaussian $s_e = 2.1$. Rounding to the nearest integer (i.e., using a continuity correction), the linear regression assuming Gaussian residuals implies that the probability for a negative number of tornados is $\Phi[(–0.5 – 3.92)/2.1] = \Phi[–2.10] = 0.018$, rather than the true value of zero. On the other hand, conditional on a temperature of 70°F, the Poisson regression specifies that the number of tornados will be distributed as a Poisson variable with mean $\mu = 3.82$. Using this mean, Equation 4.11 yields $\Pr\{Y < 0\} = 0$, $\Pr\{Y = 0\} = 0.022$, $\Pr\{Y = 1\} = 0.084$, $\Pr\{Y=2\} = 0.160$, and so on. ◇

7.4. PREDICTOR SELECTION

7.4.1. Why Is Careful Predictor Selection Important?

Almost always there are more potential predictors available than can be used in a statistical prediction procedure, and finding good subsets of these in particular cases is more difficult than might at first be imagined. The process is definitely not as simple as adding members of a list of potential predictors until an apparently good relationship is achieved. Perhaps surprisingly, there are dangers associated with including too many predictor variables in a forecast equation.

Example 7.6. An Overfit Regression

To illustrate the dangers of too many predictors, Table 7.6 shows total winter snowfall at Ithaca (inches) for the seven winters beginning in 1980 through 1986 and four potential predictors arbitrarily taken from an almanac (Hoffman, 1988): the U.S. federal deficit (in billions of dollars), the number of personnel in the U.S. Air Force, the sheep population of the U.S. (in thousands), and the average Scholastic Aptitude Test (SAT) scores of college-bound high-school students. Obviously these are nonsense predictors, which bear no real relationship to the amount of snowfall at Ithaca.

Regardless of their lack of relevance, we can blindly offer these predictors to a computer regression package, and it will produce a regression equation. For reasons that will be made clear shortly, assume that the regression will be fit using only the six winters beginning in 1980 through 1985. That portion of available data used to produce the forecast equation is known as the *developmental sample*,

TABLE 7.6 A small data set illustrating the dangers of overfitting. Nonclimatological data were taken from Hoffman (1988).

Winter *Beginning*	Ithaca *Snowfall (in.)*	U.S. Federal *Deficit (\$ x10^9)*	U.S. Air Force *Personnel*	U.S. *Sheep (x10^3)*	Average *SAT Scores*
1980	52.3	59.6	557969	12699	992
1981	64.9	57.9	570302	12947	994
1982	50.2	110.6	582845	12997	989
1983	74.2	196.4	592044	12140	963
1984	49.5	175.3	597125	11487	965
1985	64.7	211.9	601515	10443	977
1986	65.6	220.7	606500	9932	1001

dependent sample, or *training sample*. For the developmental sample of 1980–1985, the resulting equation is

$$Snow = 1161771 - 601.7\,yr - 1.733\,deficit + 0.0567\,AF\,pers. - 0.3799\,sheep + 2.882\,SAT$$

The ANOVA table accompanying this equation indicated MSE = 0.0000, $R^2 = 100.00\%$, and $F = \infty$; that is, a perfect fit!

Figure 7.14 shows a plot of the regression-specified snowfall totals (line segments) and the observed data (circles). For the developmental portion of the record, the regression does indeed represent the data exactly, as indicated by the ANOVA statistics, even though it is obvious from the nature of the predictor variables that the specified relationship is not physically meaningful. In fact, essentially any five predictors would have produced exactly the same perfect fit (although with different regression coefficients, b_k) to the six developmental data points. More generally, any $K = n - 1$ predictors will produce a perfect regression fit to any predictand for which there are n observations. This concept is easiest to see for the case of $n = 2$, where a straight line can be fit using any $K = 1$ predictor (simple linear regression), since a line can be found that will pass through any two points in the plane, and only an intercept and a slope are necessary to define a line. The problem, however, generalizes to any sample size.

This example illustrates an extreme case of *overfitting* the data. That is, so many predictors have been used that an excellent fit has been achieved on the dependent data, but the fitted relationship falls apart when used with independent, or *verification data*—data not used in the development of the equation. Here the data for 1986 have been reserved as a verification sample. Figure 7.14 indicates that the equation performs very poorly outside of the training sample, producing a meaningless forecast for negative snowfall during 1986–1987. Clearly, issuing forecasts equal to the climatological average total snowfall, or the snowfall for the previous winter, would yield better results than this overfit regression equation. Note that the problem of overfitting is *not* limited to cases where nonsense predictors are used in a forecast equation, and will be a problem when too many meaningful predictors are included as well. ◇

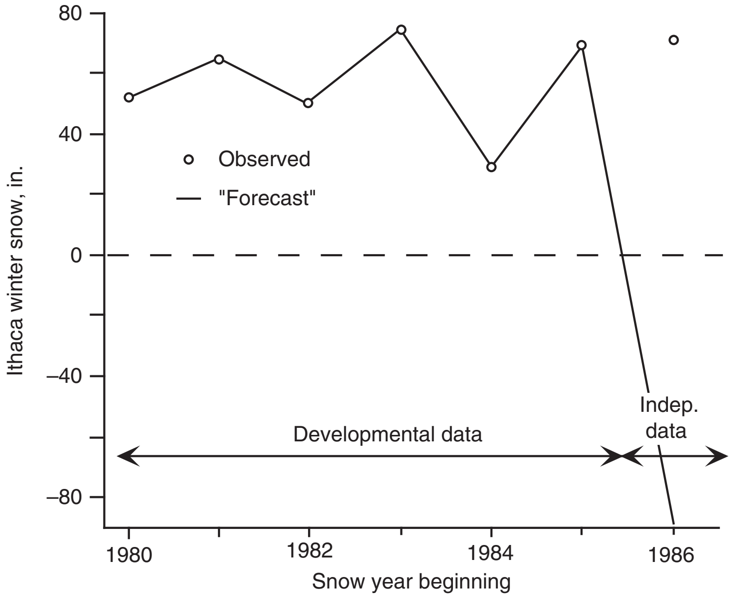


FIGURE 7.14 Forecasting Ithaca winter snowfall using the data in Table 7.6. The number of predictors is one fewer than the number of observations of the predictand in the developmental data, yielding perfect correspondence between the values specified by the regression and the predictand data for this portion of the record. The relationship falls apart completely when used with the 1986 data, which was not used in equation development. The regression equation has been grossly overfit.

As ridiculous as it may seem, several important lessons can be drawn from Example 7.6:

- Begin development of a regression equation by choosing only physically reasonable or meaningful potential predictors. If the predictand of interest is surface temperature, for example, then temperature-related predictors such as the 1000–700 mb thickness (reflecting the mean virtual temperature in the layer), the 700 mb relative humidity (perhaps as a proxy for clouds), or the climatological average temperature for the forecast date (as a representation of the annual cycle of temperature) could be sensible candidate predictors. Understanding that clouds will form only in saturated air, a binary variable based on the 700 mb relative humidity also might be expected to contribute meaningfully to the regression. One consequence of this lesson is that a statistically literate person with insight into the physical problem ("domain expertise") may be more successful than a statistician at devising a forecast equation.
- A tentative regression equation needs to be tested on a sample of data not involved in its development. One way to approach this important step is simply to reserve a portion (perhaps a quarter, a third, or half) of the available data as the independent verification set, and fit the regression using the remainder as the training set. The performance of the resulting equation will nearly always be better for the dependent than the independent data, since (in the case of least-squares regression) the coefficients have been chosen specifically to minimize the squared residuals in the developmental sample. A very large difference in performance between the dependent and independent samples would lead to the suspicion that the equation had been overfit.

- We need a reasonably large developmental sample if the resulting equation is to be stable. Stability is usually understood to mean that the fitted coefficients are also applicable to independent (i.e., future) data, so that the coefficients would be substantially unchanged if based on a different sample of the same kind of data. The number of coefficients that can be estimated with reasonable accuracy increases as the sample size increases, although in weather forecasting practice one often finds that there is little to be gained from including more than about a dozen predictor variables in a final regression equation (Glahn, 1985). In that kind of forecasting application there are typically thousands of observations of the predictand in the developmental sample. Unfortunately, there is not a firm rule specifying a minimum ratio of sample size (number of observations of the predictand) to the number of predictor variables in a final equation. Rather, testing on an independent data set is relied upon in practice to ensure stability of the regression.

7.4.2. Screening Predictors

Suppose the set of potential predictor variables for a particular problem could be assembled in a way that all physically relevant predictors were included, with exclusion of all irrelevant ones. This ideal can rarely, if ever, be achieved. Even if it could be, however, it generally would not be useful to include all the potential predictors in a final equation. This is because the predictor variables are almost always mutually correlated, so that the full set of potential predictors contains redundant information. Table 3.5, for example, shows substantial correlations among the six variables in Table A.1. Inclusion of predictors with strong mutual correlation is worse than superfluous because this condition leads to poor estimates (high-variance sampling distributions) for the regression parameters. As a practical matter, then, we need a method to choose among potential predictors and to decide how many and which of them are sufficient to produce a good prediction equation.

In the jargon of statistical weather forecasting, the problem of selecting a good set of predictors from a pool of potential predictors is called *screening regression*, since the potential predictors must be subjected to some kind of screening, or filtering procedure. The most commonly used screening procedure is known as *forward selection* or *stepwise regression* in the broader statistical literature.

Suppose there are some number, M, of candidate potential predictors for a least-squares linear regression. We begin the process of forward selection with the uninformative prediction equation $\hat{y} = b_0$. That is, only the intercept term is "in the equation," and this intercept is necessarily the sample mean of the predictand. On the first forward selection step, all M potential predictors are examined for the strength of their linear relationship to the predictand. In effect, all the possible M simple linear regressions between the available predictors and the predictand are computed, and that predictor whose linear regression is best among all candidate predictors is chosen as x_1. At this stage of the screening procedure, then, the prediction equation is $\hat{y} = b_0 + b_1 x_1$. Note that in general the intercept b_0 no longer will be the average of the y values.

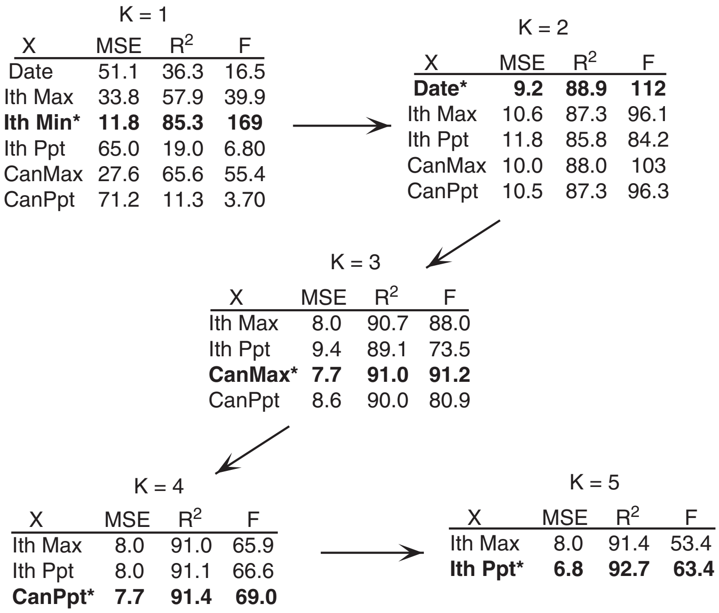
At the next stage of the forward selection, trial regressions are again constructed using all remaining $M - 1$ predictors. However, all these trial regressions also contain the variable selected on the previous step as x_1. That is, given the particular x_1 chosen on the previous step, that predictor variable yielding the best regression $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$ is chosen as x_2. This new x_2 will be recognized as best because it produces that regression equation with $K = 2$ predictors that also includes the previously chosen x_1, having the highest R^2, the smallest MSE, and the largest F ratio.

Subsequent steps in the forward selection procedure follow this pattern exactly: at each step, that member of the potential predictor pool not yet in the regression is chosen that produces the best regression in conjunction with the $K - 1$ predictors chosen on previous steps. In general, when these regression equations are recomputed, the regression coefficients for the intercept and for the previously chosen predictors will change. These changes will occur because the predictors usually are correlated to a greater or lesser degree, so that information about the predictand is spread among the predictors differently as more predictors are added to the equation.

Example 7.7. Equation Development Using Forward Selection

The concept of predictor selection can be illustrated with the January 1987 temperature and precipitation data in Table A.1. As in Example 7.1 for simple linear regression, the predictand is Canandaigua minimum temperature. The potential predictor pool consists of maximum and minimum temperatures at Ithaca, maximum temperature at Canandaigua, the logarithms of the precipitation amounts plus 0.01 in. (in order for the logarithm to be defined for zero precipitation) for both locations, and the day of the month. The date predictor is included on the basis of the trend in the residuals apparent in Figure 7.9. Note that this example is somewhat artificial with respect to statistical weather forecasting, since the predictors (other than the date) will not be known in advance of the time that the predictand (minimum temperature at Canandaigua) will be observed. However, this small data set serves perfectly well to illustrate the principles.

Figure 7.15 diagrams the process of choosing predictors using forward selection. The numbers in each table summarize the comparisons being made at each step. For the first ($K = 1$) step, no predictors are yet in the equation, and all six potential predictors are under consideration. At this stage the



K = 1

X	MSE	R^2	F
Date	51.1	36.3	16.5
Ith Max	33.8	57.9	39.9
Ith Min*	**11.8**	**85.3**	**169**
Ith Ppt	65.0	19.0	6.80
CanMax	27.6	65.6	55.4
CanPpt	71.2	11.3	3.70

K = 2

X	MSE	R^2	F
Date*	**9.2**	**88.9**	**112**
Ith Max	10.6	87.3	96.1
Ith Ppt	11.8	85.8	84.2
CanMax	10.0	88.0	103
CanPpt	10.5	87.3	96.3

K = 3

X	MSE	R^2	F
Ith Max	8.0	90.7	88.0
Ith Ppt	9.4	89.1	73.5
CanMax*	**7.7**	**91.0**	**91.2**
CanPpt	8.6	90.0	80.9

K = 4

X	MSE	R^2	F
Ith Max	8.0	91.0	65.9
Ith Ppt	8.0	91.1	66.6
CanPpt*	**7.7**	**91.4**	**69.0**

K = 5

X	MSE	R^2	F
Ith Max	8.0	91.4	53.4
Ith Ppt*	**6.8**	**92.7**	**63.4**

FIGURE 7.15 Diagram of the forward selection procedure for development of a regression equation for Canandaigua minimum temperature using as potential predictors the remaining variables in data set A.1, plus the date. At each step the variable is chosen (bold, starred) whose addition would produce the largest decrease in MSE or, equivalently, the largest increase in R^2 or F. At the final ($K = 6$) stage, only Ith. Max remains to be chosen, and its inclusion would produce MSE = 6.8, $R^2 = 93.0\%$, and $F = 52.8$.

predictor producing the best simple linear regression is chosen, as indicated by the smallest MSE, and the largest R^2 and F ratio among the six. This best predictor is the Ithaca minimum temperature, so the tentative regression equation is exactly Equation 7.20.

Having chosen the Ithaca minimum temperature in the first stage, there are five potential predictors remaining, and these are listed in the $K = 2$ table. Of these five, the one producing the best predictions in an equation that also includes the Ithaca minimum temperature is chosen. Summary statistics for these five possible two-predictor regressions are also shown in the $K = 2$ table. Of these, the equation including Ithaca minimum temperature and the date as the two predictors is clearly best, producing MSE $= 9.2°\text{F}^2$ for the dependent data.

With these two predictors now in the equation, there are only four potential predictors left at the $K = 3$ stage. Of these, the Canandaigua maximum temperature produces the best predictions in conjunction with the two predictors already in the equation, yielding MSE $= 7.7°\text{F}^2$ for the dependent data. Similarly, the best predictor at the $K = 4$ stage is Canandaigua precipitation, and the better predictor at the $K = 5$ stage is Ithaca precipitation. For $K = 6$ (all predictors in the equation), the MSE for the dependent data is $6.8°\text{F}^2$, with $R^2 = 93.0\%$. ◇

An alternative approach to screening regression is called *backward elimination*. The process of backward elimination is analogous but opposite to that of forward selection. Here the initial stage is a regression containing all M potential predictors, $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_Mx_M$, so backward elimination will not be computationally feasible if $M \geq n$. Usually this initial equation will be grossly overfit, containing many redundant and some possibly useless predictors. At each step of the backward elimination procedure, the least important predictor variable is removed from the regression equation. That variable will be the one whose coefficient is smallest in absolute value, relative to its estimated standard error. In terms of the sample regression output tables presented earlier, the removed variable will exhibit the smallest (absolute) t ratio. As in forward selection, the regression coefficients for the remaining variables require recomputation if (as is usually the case) the predictors are mutually correlated.

There is no guarantee that forward selection and backward elimination will choose the same subset of the potential predictor pool for the final regression equation. Other predictor selection procedures for multiple regression also exist, and these might select still different subsets. The possibility that a chosen selection procedure might not select the "right" set of predictor variables might be unsettling at first, but as a practical matter this is not usually an important problem in the context of producing an equation for use as a forecast tool. Correlations among the predictor variables often result in the situation that essentially the same information about the predictand can be extracted from different subsets of the potential predictors. Therefore, if the aim of the regression analysis is only to produce reasonably accurate forecasts of the predictand, the black box approach of empirically choosing a workable set of predictors is quite adequate. However, we should not be so complacent in a research setting, where one aim of a regression analysis could be to find specific predictor variables most directly responsible for the physical phenomena associated with the predictand.

7.4.3. Stopping Rules

Both forward selection and backward elimination require a stopping criterion, or stopping rule. Without such a rule, forward selection would continue until all M candidate predictor variables were included in the regression equation, and backward elimination would continue until all predictors had been eliminated. It might seem that finding the stopping point would be a simple matter of

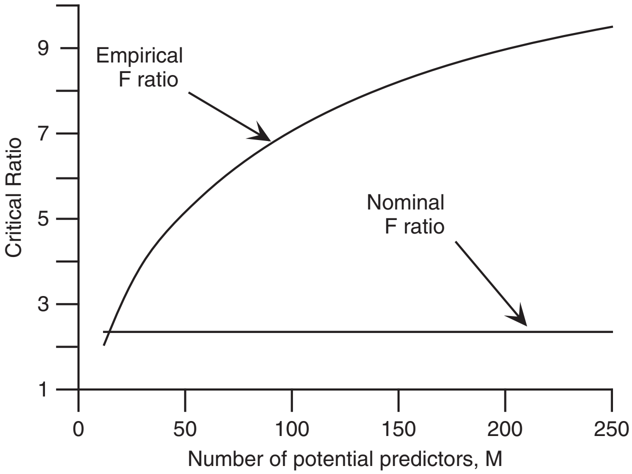


FIGURE 7.16 Comparison of the nominal and empirically (resampling-) estimated critical ($p = 0.01$) F ratios for overall significance in a particular regression problem, as a function of the number of potential predictor variables, M. The sample size is $n = 127$, with the best $K = 12$ predictor variables to be included in each final regression equation. The nominal F ratio of 2.35 is applicable only for the case of $M = K$. When the forward selection procedure can choose from among more than K potential predictors the true critical F ratio is substantially higher. The difference between the nominal and actual values widens as M increases. *From Neumann* et al. *(1977).*

evaluating the test statistics for the regression parameters and their nominal p values as supplied by the computer regression software. Unfortunately, because of the way the predictors are selected, these implied hypothesis tests are not quantitatively applicable. At each step (either in selection or elimination) predictor variables are not chosen randomly for entry or removal. Rather, the best or worst, respectively, among the available choices is selected. Although this may seem like a minor distinction, it can have very major consequences.

The problem is illustrated in Figure 7.16, taken from the study of Neumann et al. (1977). The specific problem represented in this figure is the selection of exactly $K = 12$ predictor variables from pools of potential predictors of varying sizes, M, when there are $n = 127$ observations of the predictand. Ignoring the problem of nonrandom predictor selection would lead us to declare as significant any regression for which the F ratio in the ANOVA table is larger than the nominal critical value of 2.35. Naïvely, this value would correspond to the minimum F ratio necessary to reject the null hypothesis of no real relationship between the predictand and the 12 predictors at the 1% level. The curve labeled empirical F ratio was arrived at using a resampling test, in which the same meteorological predictor variables were used in a forward selection procedure to predict 100 artificial data sets of $n = 127$ independent Gaussian random numbers each. This procedure simulates a situation consistent with the null hypothesis that the predictors bear no real relationship to the predictand, while automatically preserving the correlations among this particular set of predictors.

Figure 7.16 indicates that the nominal regression diagnostics give the correct answer only in the case of $K = M$, for which there is no ambiguity in the predictor selection since all the $M = 12$ potential predictors must be used to construct the $K = 12$ predictor equation. When the forward selection procedure has available some larger number $M > K$ potential predictor variables to choose from, the true critical F ratio is higher, and sometimes by a substantial amount. Even though none of the potential predictors in the resampling procedure bears any real relationship to the artificial (random) predictand, the forward selection procedure chooses those predictors exhibiting the highest chance correlations with the predictand, and these relationships result in apparently large *F* ratio statistics. Put another way, the *p* value associated with the nominal critical $F = 2.35$ is too large (less significant), by an amount that increases as more potential predictors are offered to the forward selection procedure. To emphasize the seriousness of the problem, the nominal *F* ratio in the situation of Figure 7.16 for the very stringent 0.01% level test is only about 3.7. The practical result of relying literally on the nominal critical *F* ratio is to allow more predictors into the final equation than are meaningful, with the danger that the regression will be overfit. The *F* ratio in Figure 7.16 is a single-number regression diagnostic convenient for illustrating the effects of overfitting, but these effects would be reflected in other aspects of the ANOVA table also. For example, most if not all of the nominal *t* ratios for the individual cherry-picked predictors when $M >> K$ would be larger than 2 in absolute value, incorrectly suggesting meaningful relationships with the (random) predictand.

Unfortunately, the results in Figure 7.16 apply only to the specific data set from which they were derived. In order to employ this approach to estimate the true critical *F*-ratio using resampling methods, it must be repeated for each regression to be fit, since the statistical relationships among the potential predictor variables will be different in different data sets. In practice, other less rigorous stopping criteria usually are employed. For example, we might stop adding predictors in a forward selection when none of the remaining predictors would reduce the R^2 by a specified amount, perhaps 0.05%.

The stopping criterion can also be based on the MSE. This choice is intuitively appealing because, as the standard deviation of the residuals around the regression function, √MSE directly reflects the anticipated precision of a regression. For example, if a regression equation were being developed to forecast surface temperature, little would be gained by adding more predictors if the MSE were already $0.01°F^2$, since this would indicate a $\pm 2s_e$ (i.e., approximately 95%) prediction interval around the forecast value of about $\pm\ 2\ \sqrt{0.01°F^2} = 0.2\ °F$. As long as the number of predictors *K* is substantially less than the sample size *n*, adding more predictor variables (even meaningless ones) will decrease the MSE for the developmental sample. This concept is illustrated schematically in

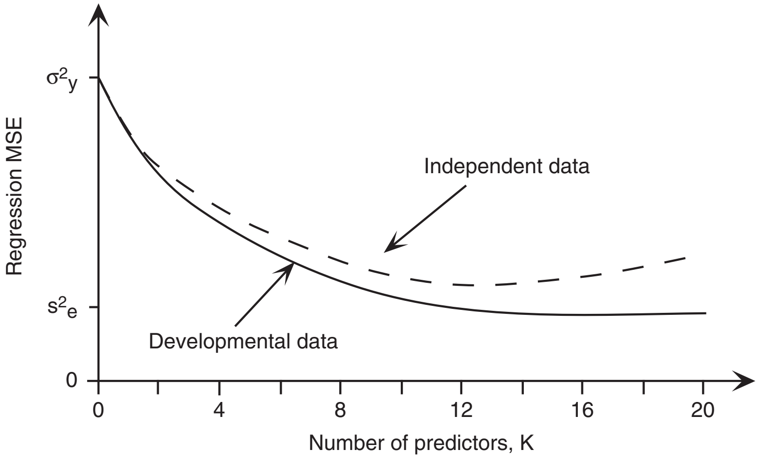


FIGURE 7.17 Schematic illustration of the regression MSE as a function of the number of predictor variables in the equation, *K*, for developmental data (solid) and for an independent verification set (dashed). After Glahn (1985).

Figure 7.17. Ideally, the stopping criterion would be activated at the point where the MSE does not decline appreciably with the addition of more predictors, at perhaps $K = 12$ predictors in the hypothetical case shown in Figure 7.17.

Figure 7.17 indicates that the MSE for an independent data set will be larger than that achieved for the developmental data. This result should not be surprising, since the least-squares fitting procedure operates by optimizing the parameter values to minimize MSE for the developmental data. This underestimation of the independent-data MSE provided by the MSE for a forecast equation on developmental data is an expression of what is sometimes called *artificial skill* (Davis, 1976; Michaelson, 1987). The precise magnitude of the differences in MSE between developmental and independent data sets is not determinable solely from the regression output using the developmental data. That is, having seen only the regressions fit to the developmental data, we cannot know the value of the minimum MSE for independent data. Neither can we know if it will occur at a similar point (at around $K = 12$ in Figure 7.17), or whether the equation has been overfit and the minimum MSE for the independent data will be for a substantially smaller K. This situation is unfortunate because the purpose of developing a forecast equation is to specify future, unknown values of the predictand using observations of the predictors that have yet to occur.

Figure 7.17 also indicates that, for forecasting purposes, the exact stopping point is not usually critical as long as it is approximately right. Again, this is because the MSE tends to change relatively little through a range of K near the optimum, and for purposes of forecasting it is the minimization of the MSE rather than the specific identities of the predictors that is important. By contrast, if the purpose of the regression analysis is scientific understanding, the specific identities of chosen predictor variables can be critically important, and the magnitudes of the resulting regression coefficients may lead to significant physical insight. In this case it is not reduction of prediction MSE, per se, that is desired, but rather that causal relationships between particular variables be suggested by the analysis.

7.4.4. Cross Validation

Often regression equations to be used for weather forecasting are tested on a sample of independent data that has been held back during development of the forecast equation. In this way, once the number K and specific identities of the predictors have been fixed, an estimate of the distances between the solid and dashed MSE lines in Figure 7.17 can be estimated directly from the reserved data. If the deterioration in forecast precision (i.e., the unavoidable increase in MSE) is judged to be acceptable, the equation can be used operationally.

This procedure of reserving an independent verification data set is actually a special case of a technique known as *cross validation* (Efron and Gong, 1983; Efron and Tibshirani, 1993; Elsner and Schmertmann, 1994; Michaelson, 1987). Cross validation simulates prediction for future, unknown data by repeating the entire fitting procedure on data subsets, and then examining the predictions made for the data portions left out of each subset. The most frequently used procedure is known as *leave-one-out cross validation*, in which the fitting procedure is repeated n times, each time with a sample of size $n - 1$, because one of the predictand observations and its corresponding predictor set are left out in each replication of the fitting process. The result is n (often only slightly) different prediction equations.

The cross-validation estimate of the prediction MSE is computed by forecasting each omitted observation using the equation developed from the remaining $n - 1$ data values, computing the squared difference between the prediction and predictand for each of these equations, and averaging the n squared differences. Thus, leave-one-out cross validation uses all n observations of the predictand to estimate the prediction MSE in a way that allows each observation to be treated, one at a time, as independent data.

It should be emphasized that each repetition of the cross-validation exercise is a repetition of the entire fitting algorithm, not a refitting of the specific statistical model derived from the full data set, using $n - 1$ data values. In particular, different prediction variables must be allowed to enter for different cross-validation subsets. DelSole and Shukla (2009) provide a cautionary analysis showing that failure to respect this precept can lead to random-number predictors exhibiting apparently real, cross-validated predictive ability. Any data transformations (e.g., standardizations with respect to climatological values) also need to be defined (and therefore possibly recomputed) without any reference to the withheld data in order for them to have no influence on the equation that will be used to predict them in the cross-validation exercise. However, the ultimate product equation, to be used for operational forecasts, would be fit using all the data after we are satisfied with the cross-validation results.

Cross validation can also be carried out for any number m of withheld data points and developmental data sets of size $n - m$ (Zhang, 1993). In this more general case, as many as all $(n!)/[(m!)(n - m)!]$ possible partitions of the full data set could be employed. Particularly when the sample size n is small and the predictions will be evaluated using a correlation measure, leaving out $m > 1$ values at a time can be advantageous (Barnston and van den Dool, 1993).

Cross validation requires some special care when the data are serially correlated. In particular, data records adjacent to or near the omitted observation(s) will tend to be more similar to them than randomly selected ones, so the omitted observation(s) will be more easily predicted than the uncorrelated future observations they are meant to simulate. A solution to this problem is to leave out blocks of an odd number of consecutive observations, L, so the fitting procedure is repeated $n - L + 1$ times on samples of size $n - L$ (Burman et al., 1994; Elsner and Schmertmann, 1994). The blocklength L is chosen to be large enough for the correlation between its middle value and the nearest data used in the cross-validation fitting to be small, and the cross-validation prediction is made only for that middle value. For $L = 1$ this moving-blocks cross validation reduces to leave-one-out cross validation.

Another elaboration on cross validation that can be used with serially correlated data, and that may be preferable to the leave-one-out approach for large samples of uncorrelated data, is to successively leave out one of L nonoverlapping blocks of data. For example, for $L = 5$, fivefold cross validation repeats the fitting exercise five times, each with 20% of the data reserved for verification. $L = n$ yields the leave-one-out procedure. Hastie et al. (2009) suggest use of $L = 5$ or 10.

Example 7.8. Protecting against Overfitting Using Cross Validation

Having used all the available developmental data to fit the regressions in Example 7.7, what can be done to ensure that these prediction equations have not been overfit? Fundamentally, what is desired is a measure of how the regressions will perform when used on data not involved in the fitting. Cross validation is an especially appropriate tool for this purpose in the present example because the small ($n = 31$) sample would be inadequate if a substantial portion of it had to be reserved for a validation sample.

Figure 7.18 evaluates MSEs for six regression equations obtained with forward selection. This figure shows real results in the same form as the idealization of Figure 7.17. The solid line indicates the MSE achieved on the developmental sample, obtained by adding the predictors in the order shown in Figure 7.15. Because a regression chooses precisely those coefficients minimizing MSE for the developmental data, this quantity is expected to be higher when the equations are applied to independent data. An estimate of how much higher is given by the MSEs from the cross-validation samples (dashed line). Because these data are autocorrelated, a simple leave-one-out cross validation is expected to underestimate the prediction MSE. Here the cross validation has been carried out omitting

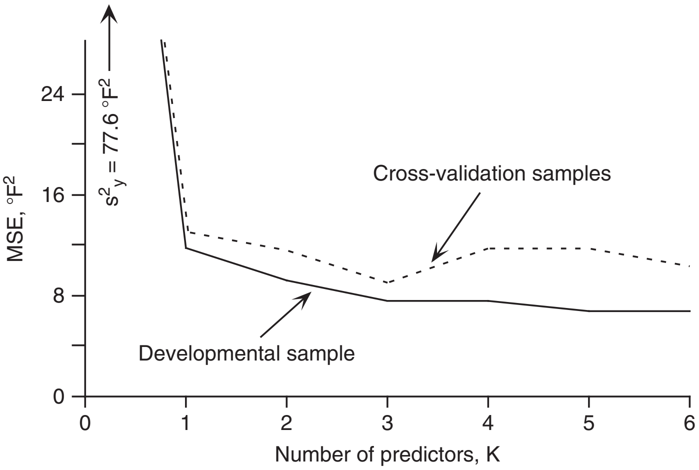


FIGURE 7.18 Plot of residual mean-squared error as a function of the number of regression predictors specifying Canandaigua minimum temperature, using the January 1987 data in Appendix A. Solid line shows MSE for developmental data (starred predictors in Figure 7.15). Dashed line shows MSE achievable on independent data, with the same numbers of (possibly different) predictors, as estimated through cross validation, leaving out blocks of seven consecutive days. This plot is a real-data example corresponding to the idealization in Figure 7.17.

blocks of length $L = 7$ consecutive days. Since the lag-1 autocorrelation for the predictand is approximately $r_1 = 0.6$ and the autocorrelation function exhibits approximately exponential decay (similar to that in Figure 3.20), the correlation between the predictand in the centers of the seven-day moving blocks and the nearest data used for equation fitting is $0.6^4 = 0.13$, corresponding to $R^2 = 1.7\%$, indicating near-independence.

Each cross-validation point in Figure 7.18 represents the average of 25 ($= 31 - 7 + 1$) squared differences between an observed value of the predictand at the center of a block and the forecast of that value produced by regression equations fit to all the data except those in that block. Predictors are added to each of these equations according to the usual forward selection algorithm. The order in which the predictors are added in one of these 25 regressions is often the same as that indicated in Figure 7.15 for the full data set, but this order is not forced onto the cross-validation samples; indeed it is different for some of the data partitions.

The differences between the dashed and solid lines in Figure 7.18 are indicative of the expected prediction errors for future independent data (dashed), and those that would be inferred from the MSE on the dependent data as provided by the ANOVA table (solid). The minimum cross-validation MSE at $K = 3$ suggests that the best regression for these data may be the one with three predictors, and that it should produce prediction MSE on independent data of around 9.1°F^2, yielding $\pm 2 s_e$ confidence limits of ±6.0°F. ◇

Before leaving the topic of cross validation, it is worthwhile to note that the procedure is sometimes mistakenly referred to as the *jackknife*, a relatively simple resampling procedure that was introduced in Section 5.3.5. The confusion is understandable because the jackknife is computationally analogous to leave-one-out cross validation. Its purpose, however, is to estimate the bias and/or standard deviation of a sampling distribution nonparametrically, and using only the data in a single sample. Given a sample of n independent observations, the idea in jackknifing is to recompute a statistic of interest n times, omitting a different one of the data values each time. Attributes of the sampling distribution for the statistic can then be inferred from the resulting n-member jackknife distribution (Efron 1982; Efron and Tibshirani 1993). The jackknife and leave-one-out cross validation share the mechanics of repeated recomputation on reduced samples of size $n - 1$, but cross validation seeks to infer future forecasting performance, whereas the jackknife seeks to nonparametrically characterize the sampling distribution of a sample statistic.

7.5. OBJECTIVE FORECASTS USING TRADITIONAL STATISTICAL METHODS

7.5.1. Classical Statistical Forecasting

Construction of weather forecasts through purely statistical means—that is, without the benefit of information from fluid-dynamical weather prediction models—has come to be known as classical statistical forecasting. This name reflects the long history of the use of purely statistical forecasting methods, dating from the time before the availability of dynamical forecast information. The accuracy of dynamical forecasts has advanced sufficiently that pure statistical forecasting is used in practical settings only for very short lead times or for fairly long lead times.

Very often classical forecast products are based on multiple regression equations of the kinds described in Sections 7.2 and 7.3. These statistical forecasts are objective in the sense that a particular set of inputs or predictors will always produce the same forecast for the predictand, once the forecast equation has been developed. However, many subjective decisions necessarily go into the development of the forecast equations.

The construction of a classical statistical forecasting procedure follows from a straightforward implementation of the ideas presented in the previous sections of this chapter. Required developmental data consist of past values of the quantity to be forecast and a matching collection of potential predictors whose values will be known prior to the forecast time. A forecasting procedure is developed using this set of historical data, which can then be used to forecast future values of the predictand on the basis of future observations of the predictor variables. It is thus a characteristic of classical statistical weather forecasting that the time lag is built directly into the forecast equation through the time-lagged relationships between the predictors and the predictand.

For lead times up to a few hours, purely statistical forecasts still find productive use. This short-lead forecasting niche is known as *nowcasting*. Dynamically based forecasts are not practical for nowcasting because of the delays introduced by the processes of gathering weather observations, data assimilation (calculation of initial conditions for the dynamical model), the actual running of the forecast model, and the postprocessing and dissemination of the results. One very simple statistical approach that can produce competitive nowcasts is use of *conditional climatology*—that is, historical statistics subsequent to (conditional on) analogous weather situations in the past. The result could be a conditional frequency distribution for the predictand, or a single-valued forecast corresponding to the expected value (mean) of that conditional distribution. A more sophisticated approach is to construct a regression equation to forecast a few hours ahead. For example, Vislocky and Fritsch (1997) compare these two approaches for forecasting airport ceiling and visibility at lead times of one, three, and six hours.

At lead times beyond perhaps 10 days to two weeks, statistical forecasts are again competitive with dynamical forecasts. At these longer lead times, the sensitivity of dynamical models to the unavoidable small errors in their initial conditions, described in Section 7.6, makes explicit forecasting of specific weather events problematic. Although long-lead forecasts for seasonally averaged quantities currently are made using dynamical models (e.g., Barnston et al., 2003), comparable or even better predictive accuracy at substantially lower cost is still obtained through statistical methods (Anderson et al.,1999; Barnston et al., 1999; Hastenrath et al., 2009; Landsea and Knaff, 2000; Moura and Hastenrath, 2004; Quan et al., 2006; van den Dool, 2007; Zheng et al., 2008). Often the predictands in these seasonal forecasts are spatial patterns, and so the forecasts involve multivariate statistical methods that are more elaborate than those described in Sections 7.2 and 7.3 (e.g., Barnston, 1994; Mason and Mimmack, 2002; Ward and Folland, 1991; see Sections 13.2.3 and 14.4). However, regression methods are still appropriate and useful for single-valued predictands. For example, Knaff and Landsea (1997) used ordinary least-squares regression for seasonal forecasts of tropical sea-surface temperatures with observed sea-surface temperatures as predictors, and Elsner and Schmertmann (1993) used Poisson regression for seasonal prediction of hurricane numbers.

Example 7.9. A Set of Classical Statistical Forecast Equations

The flavor of classical statistical forecast methods can be appreciated by looking at the NHC-67 procedure for forecasting hurricane movement (Miller et al., 1968). This relatively simple set of regression equations was used as part of the operational suite of forecast models at the U.S. National Hurricane Center until 1988 (Sheets, 1990). Since hurricane movement is a vector quantity, each forecast consists of two equations: one for northward movement and one for westward movement. The two-dimensional forecast displacement is then computed as the vector sum of the northward and westward forecasts.

The predictands were stratified according to two geographical regions: north and south of 27.5°N latitude. That is, separate forecast equations were developed to predict storms on either side of this latitude, on the basis of the subjective experience of the developers regarding the responses of hurricane movement to the larger-scale flow, and in particular on the basis that storms moving in the trade winds in the lower latitudes tend to behave less erratically. Separate forecast equations were also developed for "slow" versus "fast" storms. The choice of these two stratifications was also made subjectively, on the basis of the experience of the developers. Separate equations are also needed for each forecast lead time (0 – 12h, 12 – 24h, 24 – 36h, and 36 – 48h, yielding a total of 2 (displacement directions) x 2 (regions) x 2 (speeds) x 4 (lead times) = 32 separate regression equations in the NHC-67 package.

The available developmental data set consisted of 236 northern cases (initial position for hurricanes) and 224 southern cases. Candidate predictor variables were derived primarily from 1000-, 700-, and 500-mb heights at each of 120 gridpoints in a 5° x 5° coordinate system that follows the storm. Predictors derived from these 3 x 120 = 360 geopotential height predictors, including 24-h height changes at each level, geostrophic winds, thermal winds, and Laplacians of the heights, were also included as candidate predictors. In addition, two persistence predictors, observed northward and westward storm displacements in the previous 12 hours, were included.

With vastly more potential predictors than observations, some screening procedure is clearly required. Here forward selection was used, with the (subjectively determined) stopping rule that no more than 15 predictors would be in any equation, and new predictors would be only included to the extent that they increased the regression R^2 by at least 1%. This second criterion was apparently sometimes relaxed for regressions with few predictors.

Table 7.7 shows the results for the 0–12h westward displacement of slow southern storms in NHC-67. The five predictors are shown in the order they were chosen by the forward selection procedure, together with the R^2 value achieved on the developmental data at each step. The coefficients are those for the final ($K = 5$) equation. The most important single predictor was the persistence variable (P_x), reflecting the tendency of hurricanes to change speed and direction fairly slowly. The 500-mb height at a point north and west of the storm (Z_{37}) corresponds physically to the steering effects of midtropospheric flow on hurricane movement. Its coefficient is positive, indicating a tendency for westward storm displacement given relatively high heights to the northwest, and slower or eastward (negative westward) displacement of storms located southwest of 500-mb troughs. The final two or three predictors appear to improve the regression only marginally—the predictor Z_3 increases the R^2 by less than 1%—and it is quite possible that the $K = 2$ or $K = 3$ predictor models might have been chosen, and might have been equally accurate for independent data, if cross validation had been computationally feasible for the developers. Remarks in Neumann et al. (1977) concerning the fitting of the similar NHC-72 regressions, in relation to Figure 7.16, are also consistent with the idea that the equation represented in Table 7.7 may have been overfit. ◇

7.5.2. Perfect Prog and MOS

Pure classical statistical weather forecasts for lead times in the range of a few days are generally no longer employed, since dynamical models now allow more accurate forecasts at this timescale. However, two types of statistical weather forecasting are in use that improve on aspects of dynamical forecasts, essentially by postprocessing their raw output. Both of these methods use large multiple regression equations in a way that is analogous to the classical approach, so that many of the same

TABLE 7.7 Regression results for the NHC-67 hurricane forecast procedure, for the 0–12h westward displacement of slow southern zone storms, indicating the order in which the predictors were selected and the resulting R^2 at each step. The meanings of the symbols for the predictors are P_X = westward displacement in the previous 12 h, Z_{37} = 500-mb height at the point 10° north and 5° west of the storm, P_Y = northward displacement in the previous 12 h, Z_3 = 500-mb height at the point 20° north and 20° west of the storm, and P_{51} = 1000-mb height at the point 5° north and 5° west of the storm. Distances are in nautical miles, and heights are in meters. From Miller et al. (1968).

Predictor	Coefficient	Cumulative R^2
Intercept	–2709.5	—
P_x	0.8155	79.8%
Z_{37}	0.5766	83.7%
P_y	–0.2439	84.8%
Z_3	–0.1082	85.6%
P_{51}	–0.3359	86.7%

technical considerations pertaining to equation fitting apply. The differences between these two approaches and classical statistical forecasting have to do with the range of available predictor variables. In addition to conventional predictors such as current meteorological observations, the date, or climatological values of a particular meteorological element, predictor variables taken from the outputs of the dynamical models are also used.

There are three reasons why statistical reinterpretation of dynamical forecast output is useful for practical weather forecasting:

- There are important differences between the real world and its representation in the dynamical models, and these differences have important implications for the forecast enterprise (Gober et al., 2008). Figure 7.19 illustrates some of these differences. Dynamical models necessarily simplify and homogenize surface conditions, by representing the world as an array of gridpoints to which the forecast output pertains. As implied by Figure 7.19, small-scale effects (e.g., of topography or small bodies of water) important to local weather may not be included in the dynamical model. Also, locations and variables for which forecasts are needed may not be represented explicitly. However, statistical relationships can be developed between the information provided by the dynamical models and desired forecast quantities to help alleviate these problems.
- Dynamical models are not complete and true representations of the workings of the atmosphere, particularly at the smaller time and space scales, and they are inevitably initialized at states that differ from the true initial state of the atmosphere. For both of these reasons, their forecasts are subject to errors. To the extent that these errors are systematic, statistical postprocessing can compensate and correct forecast biases.
- The dynamical models are deterministic. That is, even though the future state of the weather is inherently uncertain, a single integration is capable of producing only a single forecast for any meteorological element, given a particular set of initial model conditions. Using dynamical forecast information in conjunction with statistical methods allows quantification and expression of the uncertainty associated with different forecast situations. In particular, it is possible to derive probability forecasts, using methods such as REEP or logistic regression, using predictors taken from even a single deterministic dynamical integration.

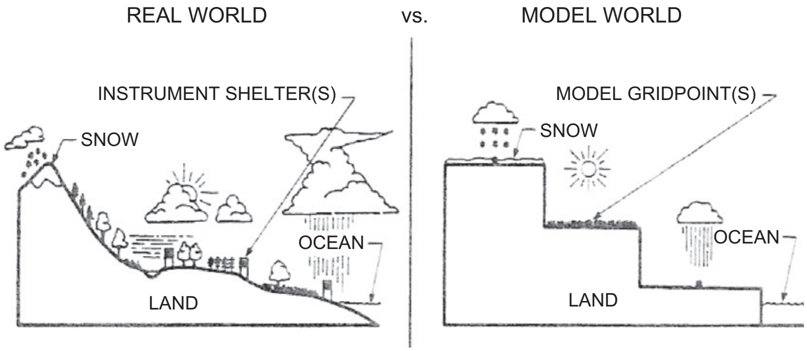


FIGURE 7.19 Cartoon illustration of differences between the real world and the world as represented by dynamical weather prediction models. *From Karl* et al. *(1989).*

The first statistical approach to be developed for taking advantage of deterministic dynamical forecasts is called "*perfect prog*" (Klein et al., 1959), which is short for perfect prognosis. As the name implies, the perfect prog technique makes no attempt to correct for possible dynamical model errors or biases, but takes the forecasts for future atmospheric variables at face value—assuming them to be perfect.

Development of perfect-prog regression equations is similar to the development of classical regression equations in that observed predictors are used to specify observed predictands. That is, only historical climatological data are used in the development of a perfect-prog forecasting equation. The primary difference between development of classical and perfect-prog equations is in the time lag. Classical equations incorporate the forecast time lag by relating predictors available before the forecast must be issued (say, today) to values of the predictand to be observed at some later time (say, tomorrow). Perfect-prog equations do not incorporate any time lag. Rather, simultaneous values of predictors and predictands are used to fit the regression equations. That is, the equations specifying tomorrow's predictand are developed using tomorrow's predictor values.

At first, it might seem that this would not be a productive approach to forecasting. Tomorrow's 1000–850 mb thickness may be an excellent predictor for tomorrow's temperature, but tomorrow's thickness will not be known until tomorrow. However, in implementing the perfect-prog approach, it is the dynamical forecasts of the predictors (e.g., today's forecast for tomorrow's thickness) that are substituted into the regression equation as predictor values. Therefore, the forecast time lag in the perfect-prog approach is contained entirely in the dynamical model. Of course quantities not forecast by the dynamical model cannot be included as potential predictors unless they will be known today. If dynamical forecasts for tomorrow's predictors really are perfect, the perfect-prog regression equations should provide very good forecasts.

The *Model Output Statistics* (MOS) approach (Carter et al., 1989; Glahn and Lowry, 1972) is the second, and usually preferred, approach to incorporating dynamical forecast information into traditional statistical weather forecasts. Preference for the MOS approach derives from its capacity to include directly in the regression equations the influences of specific characteristics of particular dynamical models at different lead times into the future.

Although both the MOS and perfect-prog approaches use quantities from dynamical integrations as predictor variables, the two approaches apply the information differently. The perfect-prog approach uses the dynamical forecast predictors only when making forecasts, but the MOS approach uses these predictors in both the development and implementation of the forecast equations. Think again in terms of today as the time at which the forecast must be made and tomorrow as the time to which the forecast pertains. MOS regression equations are developed for tomorrow's predictand using dynamical forecasts for tomorrow's values of the predictors. The true values of tomorrow's predictors are still unknown, but dynamical forecasts for them have been computed today. For example, in the MOS approach, one important predictor for tomorrow's temperature could be tomorrow's 1000–850 mb thickness as forecast today by a particular dynamical model. Therefore, to develop MOS forecast equations, it is necessary to have a developmental data set including historical records of the predictand, together with archived records of the forecasts produced by that dynamical model for the same days on which the predictand was observed.

In common with the perfect-prog approach, the time lag in MOS forecasts is incorporated through the dynamical forecast. Unlike perfect prog, the implementation of a MOS forecast equation is completely consistent with its development. That is, in both development and implementation, the MOS statistical forecast for tomorrow's predictand is made using the dynamical forecast for tomorrow's predictors, which are available today. Also unlike the perfect-prog approach, separate MOS forecast equations must be developed for different forecast lead times. This is because the error characteristics of the dynamical forecasts are different at different lead times, producing, for example, different statistical relationships between observed temperature and forecast thicknesses for 24 h versus 48 h in the future.

The classical, perfect-prog, and MOS approaches are most commonly based on multiple linear regression, exploiting correlations between a predictand and available predictors (although nonlinear regressions can also be used: e.g, Lemcke and Kruizinga, 1988; Marzban et al., 2007; Vislocky and Fritsch, 1995). In the classical approach, it is the correlations between today's values of the predictors and tomorrow's predictand that forms the basis of the forecast. For the perfect-prog approach, it is the simultaneous correlations between today's values of both predictand and predictors that are the statistical basis of the prediction equations. In the case of MOS forecasts, the prediction equations are constructed on the basis of correlations between dynamical forecasts as predictor variables and the subsequently observed value of tomorrow's predictand.

These distinctions can be expressed mathematically, as follows. In the classical approach, the forecast predictand at some future time, *t*, is expressed in the regression function f_C using a vector of (i.e., multiple) predictor variables, $\boldsymbol{x}_0$ according to

$$\hat{y}_t = f_C(\mathbf{x}_0). \tag{7.38}$$

The subscript *0* on the predictors indicates that they pertain to values observed at or before the time that the forecast must be formulated, which is earlier than the time *t* to which the forecast pertains. This equation emphasizes that the forecast time lag is built into the regression. It is applicable to both the development and implementation of a classical statistical forecast equation.

By contrast, the perfect-prog (PP) approach operates differently for development versus implementation of the forecast equation, and this distinction can be expressed as

$$\hat{y}_0 = f_{PP}(\mathbf{x}_0) \text{ in development,} \tag{7.39a}$$

and

$$\hat{y}_t = f_{PP}(\mathbf{x}_t) \text{ in implementation.} \tag{7.39b}$$

The perfect-prog regression function, f_{PP} is the same in both cases, but it is developed entirely with observed predictor data having no time lag with respect to the predictand. In implementation it operates on forecast values of the predictors for the future time *t*, as obtained from a dynamical model.

Finally, the MOS approach uses the same equation in development and implementation,

$$\hat{y}_t = f_{MOS}(\mathbf{x}_\mathrm{t}). \tag{7.40}$$

This equation is derived using the dynamical forecast predictors $\boldsymbol{x}_t$, pertaining to the future time *t* (but known at time *0* when the forecast will be issued), and is implemented in the same way. In common with the perfect-prog approach, the time lag is carried by the dynamical forecast, not the regression equation.

Since the perfect-prog and MOS approaches both draw on dynamical information, it is worthwhile to compare their advantages and disadvantages. There is nearly always a large developmental sample for perfect-prog equations, since these are fit using only historical climatological data. This is an advantage over the MOS approach, since fitting MOS equations requires an archived record of forecasts from the same dynamical model that will ultimately be used to provide input to the MOS equations.

Typically, several years of archived dynamical forecasts are required to develop a stable set of MOS forecast equations (e.g., Jacks et al., 1990). This requirement can be a substantial limitation because the dynamical models are not static. Rather, these models regularly undergo changes aimed at improving their performance. Minor changes in a dynamical model leading to reductions in the magnitudes of its random errors will not substantially degrade the performance of a set of MOS equations (e.g., Erickson et al., 1991). However, modifications to the model that change—even substantially reducing—systematic errors will require redevelopment of accompanying MOS forecast equations. Since it is a change in the dynamical model that will have necessitated the redevelopment of a set of MOS forecast equations, it is often the case that a sufficiently long developmental sample of predictors from the improved dynamical model will not be immediately available. By contrast, since the perfect-prog equations are developed using only climatological information, changes in the dynamical models should not require changes in the perfect-prog regression equations. Furthermore, improving either the random or systematic error characteristics of a dynamical model should improve the statistical forecasts produced by a perfect-prog equation.

Similarly, the same perfect-prog regression equations in principle can be used with any dynamical model or for any forecast lead time provided by a given model. Since the MOS equations are tuned to the particular error characteristics of the model for which they were developed, different MOS equations will, in general, be required for use with different dynamical models. Analogously, since the error characteristics of a dynamical model change with increasing lead time, different MOS equations are required for forecasts of the same atmospheric variable for different lead times into the future. Note, however, that potential predictors for a perfect-prog equation must be variables that are well predicted by the dynamical model with which they will be used. It may be possible to find an atmospheric predictor variable that relates closely to a predictand of interest, but that is badly forecast by a particular model. Such a variable might well be selected for inclusion in a perfect-prog equation on the basis of the relationship of its observed values to the predictand, but would be ignored in the development of a MOS equation if dynamical forecasts of that predictor bore little relationship to the predictand.

The MOS approach to statistical forecasting has two advantages over the perfect-prog approach that makes MOS the method of choice when practical. The first advantage is that model-calculated, but unobserved, quantities such as vertical velocity can be used as predictors. However, the dominating advantage of MOS over perfect prog is that systematic errors exhibited by the dynamical model are accounted for in the process of developing the MOS equations. Since the perfect-prog equations are developed without reference to the characteristics of any particular dynamical model, they cannot account for or correct their forecast errors. The MOS development procedure allows compensation for these systematic errors when the forecasts are computed. Systematic errors include such problems as progressive cooling or warming biases in the dynamical model with increasing forecast lead time, a tendency for modeled synoptic features to move too slowly or too quickly in the dynamical model, and even the unavoidable decrease in forecast accuracy at increasing lead times.

The compensation for systematic errors in a dynamical model that is accomplished by MOS forecast equations is easiest to see in relation to a simple bias in an important predictor. Figure 7.20 illustrates a hypothetical case in which surface temperature is to be forecast using the 1000–850 mb thickness. The x's in the figure represent the (unlagged, or simultaneous) relationship of a set of observed thicknesses with observed temperatures, and the circles represent the relationship between previously forecast thicknesses with the same temperature data. As drawn, the hypothetical dynamical model tends to forecast thicknesses that are too large by about 15 m. The scatter around the perfect-prog regression line (dashed) derives from the fact that there are influences on surface temperature other than those captured

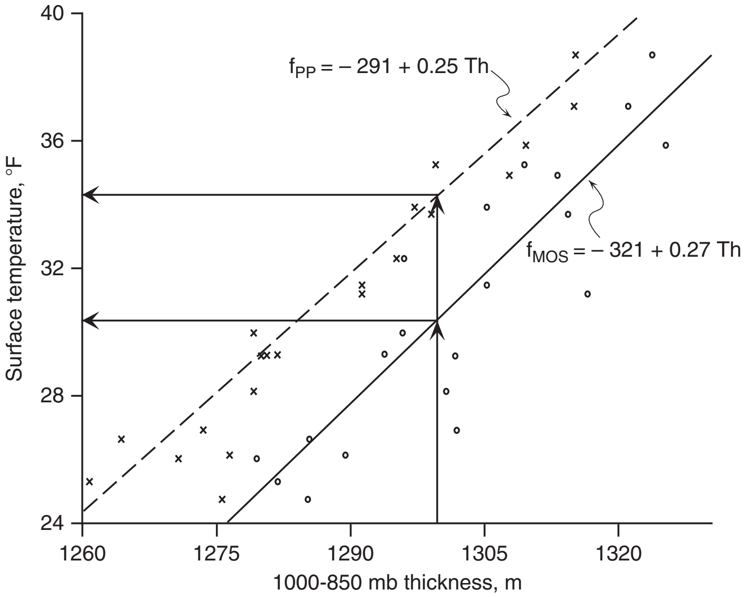


FIGURE 7.20 Illustration of the capacity of a MOS equation to correct for systematic bias in a hypothetical dynamical model. The x's represent observed, and the circles represent forecast 1000–850 mb thicknesses, in relation to hypothetical surface temperatures. The bias in the dynamical model is such that the forecast thicknesses are too large by about 15 m, on average. The MOS equation (solid line) is calibrated for this bias and produces a reasonable temperature forecast (lower horizontal arrow) when the forecast thickness is 1300 m. The perfect-prog equation (dashed line) incorporates no information regarding the attributes of the dynamical model and produces a surface temperature forecast (upper horizontal arrow) that is too warm as a consequence of the thickness bias.

by the 1000–850 mb thickness. The scatter around the MOS regression line (solid) is greater because in addition it reflects errors in the dynamical model.

The observed thicknesses (x's) in Figure 7.20 appear to specify the simultaneously observed surface temperatures reasonably well, yielding an apparently good perfect-prog regression equation (dashed line). The relationship between forecast thickness and observed temperature represented by the MOS equation (solid line) is substantially different because it includes the tendency for this dynamical model to systematically overforecast thickness. If this model produces a thickness forecast of 1300 m (vertical arrows), the MOS equation corrects for the bias in the forecast thickness and produces a reasonable temperature forecast of about 30°F (lower horizontal arrow). Loosely speaking, the MOS knows that when this dynamical model forecasts 1300 m, a more reasonable expectation for the true future thickness is closer to 1285 m, which in the climatological data (x's) corresponds to a temperature of about 30°F. The perfect-prog equation, on the other hand, operates under the assumption that a dynamical model will forecast the future thickness perfectly. It therefore yields a temperature forecast that is too warm (upper horizontal arrow) when supplied with a thickness forecast that is too large.

A more subtle systematic error exhibited by all dynamical weather forecasting models is the degradation of forecast accuracy at increasing lead time. The MOS approach accounts for this type of systematic error as well. The situation is illustrated in Figure 7.21, which is based on the hypothetical

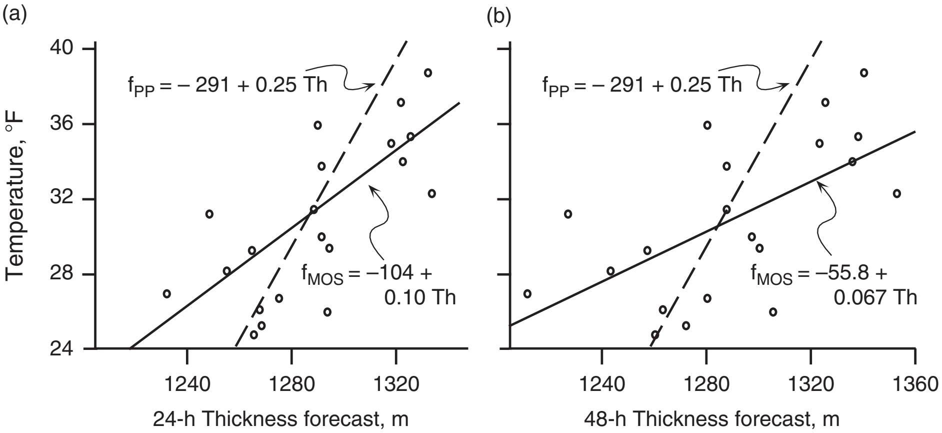


FIGURE 7.21 Illustration of the capacity of a MOS equation to account for the systematic tendency of dynamical forecasts to become less accurate at longer lead times. The points in these panels are simulated thickness forecasts, constructed from the x's in Figure 7.20 by adding random errors to the thickness values. As the forecast accuracy degrades at longer lead times, the perfect-prog equation (dashed line, reproduced from Figure 7.20) is increasingly overconfident, and tends to forecast extreme temperatures too frequently. At longer lead times (b) the MOS equations increasingly provide forecasts near the climatological average temperature (30.8°F in this example).

observed data in Figure 7.20. The panels in Figure 7.21 simulate the relationships between forecast thicknesses from an unbiased dynamical model at 24- and 48-h lead time and the surface temperature, and have been constructed by adding random errors to the observed thickness values (x's) in Figure 7.20. These random errors exhibit $\sqrt{\text{MSE}} = 20$ m for the 24-h lead time and $\sqrt{\text{MSE}} = 30$ m at the 48-h lead time. The increased scatter of points for the simulated 48-h lead time illustrates that the regression relationship is weaker when the dynamical model is less accurate.

The MOS equations (solid lines) fit to the two sets of points in Figure 7.21 reflect the progressive loss of predictive accuracy of the dynamical model at longer lead times. As the scatter of points increases, the slopes of the MOS forecast equations become more horizontal, leading to temperature forecasts that are more like the climatological mean temperature, on average. This characteristic is reasonable and desirable, since as the dynamical model provides less information about the future state of the atmosphere at longer lead times, temperature forecasts differing substantially from the climatological average temperature are progressively less well justified. In the limit of an arbitrarily long lead time, a dynamical model will really provide no more information than will the climatological value of the predictand, so that the slope of the corresponding MOS equation would be zero, and the appropriate temperature forecast consistent with this (lack of) information would simply be the climatological average temperature. Thus, it is sometimes said that MOS "converges to the climatology." By contrast, the perfect-prog equation (dashed lines, reproduced from Figure 7.20) takes no account of the decreasing accuracy of the dynamical model at longer lead times and continues to produce temperature forecasts as if the thickness forecasts were perfect. Figure 7.21 emphasizes that the result is overconfident temperature forecasts, with both very warm and very cold temperatures forecast much too frequently.

Although MOS postprocessing of dynamical forecasts is strongly preferred to perfect prog and to the raw dynamical forecasts themselves, the pace of changes made to dynamical models continues to accelerate as computing capabilities accelerate. Operationally, it would not be practical to wait for two or three years of new dynamical forecasts to accumulate before deriving a new MOS system, even if the dynamical model were to remain static for that period of time. One option for maintaining MOS systems in the face of this reality is to retrospectively *re-forecast* weather for previous years using the current updated dynamical model (Hamill et al., 2006; Jacks et al., 1990). Because daily weather data typically are strongly autocorrelated, the reforecasting process is more efficient if several days are omitted between the reforecast days (Hamill et al., 2004). Even if the computing capacity to reforecast is not available, a significant portion of the benefit of fully calibrated MOS equations can be achieved using a few months of training data (Mao et al., 1999; Neilley et al., 2002). Alternative approaches include using longer developmental data records together with whichever version of the dynamical model was current at the time and weighting the more recent forecasts more strongly. This can be done either by downweighting forecasts made with older model versions (Wilson and Valée, 2002, 2003), or by gradually downweighting older data, usually using an algorithm called the *Kalman filter* (Cheng and Steenburgh, 2007; Crochet, 2004; Galanis and Anadranistakis, 2002; Homleid, 1995; Kalnay, 2003' Mylne et al., 2002b; Valée et al., 1996), although other approaches are also possible (Yuval and Hsieh, 2003).

7.5.3. Operational MOS Forecasts

Interpretation and extension of dynamical forecasts using MOS systems has been implemented at a number of national meteorological centers, including those in the Netherlands (Lemcke and Kruizinga, 1988), Britain (Francis et al., 1982), Italy (Conte et al., 1980), China (Lu, 1991), Spain (Azcarraga and Ballester, 1991), Canada (Brunet et al., 1988), and the United States (Carter et al., 1989; Glahn et al., 2009a), among others. Most MOS applications have been oriented toward ordinary weather forecasting, but the method is equally well applicable in areas such as postprocessing of dynamical seasonal forecasts (e.g., Shongwe et al., 2006).

MOS forecast products can be quite extensive, as illustrated by Table 7.8, which shows a collection of MOS forecasts for Chicago for the 1200 UTC forecast cycle on June 14, 2010. This is one of hundreds of such panels for locations in the United States, for which these forecasts are issued twice daily and posted on the Internet by the U.S. National Weather Service. Forecasts for a wide variety of weather elements are provided, at lead times up to 60 h and at intervals as close as 3 h. After the first few lines indicating the dates and times (UTC), are forecast for daily maximum and minimum temperatures; temperatures, dew point temperatures, cloud coverage, wind speed, and wind direction at 3-h intervals; probabilities of measurable precipitation at 6- and 12-h intervals; forecasts for precipitation amount; thunderstorm probabilities; and forecast ceiling, visibility, and obstructions to visibility. Similar panels, based on several other dynamical models, are also produced and posted.

The MOS equations underlying forecasts such as those shown in Figure 7.8 are seasonally stratified, usually with separate forecast equations for the warm season (April through September) and cool season (October through March). This two-season stratification allows the MOS forecasts to incorporate different relationships between predictors and predictands at different times of the year. A finer stratification (three-month seasons, or separate month-by-month equations) would probably be preferable if sufficient developmental data were available.

The forecast equations for all elements except temperatures, dew points, and winds are regionalized. That is, developmental data from groups of nearby and climatically similar stations were composited in order to increase the sample size when deriving the forecast equations. For each regional group, then,

TABLE 7.8 Example MOS forecasts produced by the U.S. National Meteorological Center for Chicago, Illinois, shortly after 1200 UTC on June 14, 2010. A variety of weather elements are forecast, at lead times up to 60 h and at intervals as close as 3 h.

KORD	NAM		MOS			GUIDANCE								6/14/2010 1200 UTC							
DT	*/JUNE 14*		*/JUNE 15*								*/JUNE 16*								*/JUNE 17*		
HR	18	21	00	03	06	09	12	15	18	21	00	03	06	09	12	15	18	21	00	06	12
N/X							60				79				64				83		64
TMP	68	67	64	62	62	62	62	68	75	78	75	70	68	66	67	75	81	82	78	68	68
DPT	60	60	58	57	58	59	59	62	61	61	60	61	62	61	61	59	57	56	56	59	60
CLD	OV	OV	OV	OV	OV	OV	OV	OV	OV	OV	SC	BK	OV	BK	BK	SC	SC	FW	CL	CL	SC
WDR	07	05	05	04	02	23	24	27	24	26	26	22	28	26	30	32	32	35	02	20	22
WSP	08	11	12	06	02	05	05	07	10	10	08	04	03	02	04	07	08	09	08	02	02
P06			40		48		7		4		2		24		35		7		4	6	19
P12							62				6				52				10		19
Q06			1		1		0		0		0		0		0		0		0	0	0
Q12							1				0				1				0		0
T06		15/10		11/5		4/2		4/4		16/10		13/7		12/4		4/4		3/10		6/1	
T12				17/10		12/6				24/12				12/4				3/10			
CIG	2	4	5	4	2	1	2	2	4	8	8	8	6	7	7	8	8	8	8	8	8
VIS	5	5	5	5	4	4	4	5	7	7	7	7	7	7	7	7	7	7	7	7	7
OBV	BR	HZ	BR	BR	BR	BR	BR	BR	N	N	N	N	N	N	N	N	N	N	N	N	N

forecasts are made with the same equations and the same regression coefficients. This does not mean that the forecasts for all the stations in the group are the same, however, since interpolation of the dynamical output to the different forecast locations yields different predictor values. Some of the MOS equations also contain predictors representing local climatological values, which introduces further differences in the forecasts for the different stations. Regionalization is especially valuable for producing good forecasts of rare events.

In order to enhance consistency among the forecasts for different but related weather elements, some of the MOS equations are developed simultaneously. This means that the same predictor variables, though with different regression coefficients, are forced into prediction equations for related predictands in order to enhance the consistency of the forecasts. For example, it would be physically unreasonable and clearly undesirable for the forecast dew point to be higher than the forecast temperature. To help ensure that such inconsistencies appear in the forecasts as rarely as possible, the MOS equations for maximum temperature, minimum temperature, and the 3-h temperatures and dew points all contain the same predictor variables. Similarly, the four groups of forecast equations for wind speeds and directions, the 6- and 12-h precipitation probabilities, the 6- and 12-h thunderstorm probabilities, and the probabilities for precipitation types, were also developed simultaneously to enhance their consistency.

Because MOS forecasts are made for a large number of locations, it is possible to view them as maps, which are also posted on the Internet. Some of these maps display selected quantities from the MOS panels such as the one shown in Table 7.8. Figure 7.22 shows a forecast map for a predictand

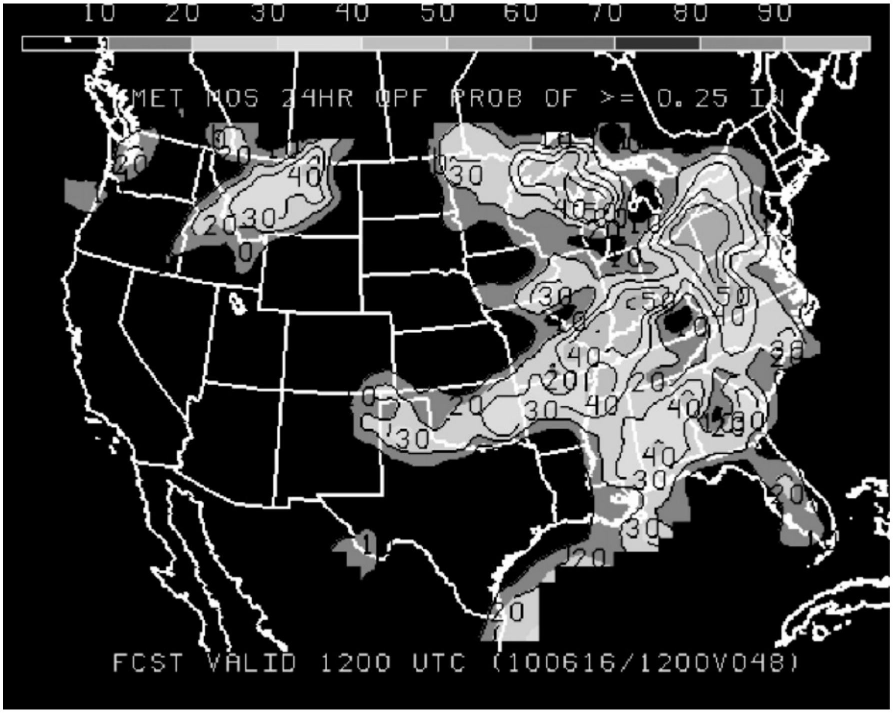

10 20 30 40 50 60 70 80 90
MET MOS 24HR QPF PROB OF >= 0.25 IN
FCST VALID 1200 UTC (100616/1200V048)

FIGURE 7.22 Example MOS forecasts in map form. The predictand is the probability of at least 0.25 in. of precipitation during a 24-h period. The contour interval is 0.10. *From www.nws.noaa.gov/mdl.*

not currently included in the tabular forecast products: probabilities of at least 0.25 in. of (liquid-equivalent) precipitation, accumulated over a 24-h period.

7.6. ENSEMBLE FORECASTING

7.6.1. Probabilistic Field Forecasts

In Section 1.3 it was asserted that dynamical chaos ensures that the future behavior of the atmosphere cannot be known with certainty. Because the atmosphere can never be fully observed, either in terms of spatial coverage or accuracy of measurements, a fluid-dynamical model of its behavior will always begin calculating forecasts from a state at least slightly different from that of the real atmosphere. These models (and other nonlinear dynamical systems, including the real atmosphere) exhibit the property that solutions (forecasts) started from only slightly different initial conditions will yield quite different results for lead times sufficiently far into the future. For synoptic-scale weather predictions, "sufficiently far" is a matter of days or (at most) weeks, and for mesoscale forecasts this window is even shorter, so that the problem of sensitivity to initial conditions is of practical importance.

Dynamical forecast models are the mainstay of weather forecasting, and the inherent uncertainty of their results must be appreciated and quantified if their information is to be utilized most effectively. For example, a single deterministic forecast of the hemispheric 500-mb height field two days in the future is at best only one member of an essentially infinite collection of 500-mb height fields that could plausibly occur. Even if this deterministic forecast of the 500-mb height field is the best possible single forecast that can be constructed, its usefulness and value will be enhanced if aspects of the probability distribution of which it is a member can be estimated and communicated. This is the problem of probabilistic field forecasting.

Probability forecasts for scalar quantities, such as a maximum daily temperature at a single location, are relatively straightforward. Many aspects of producing such forecasts have been discussed in this chapter, and the uncertainty of such forecasts can be expressed using univariate probability distributions of the kind described in Chapter 4. However, producing a probability forecast for a field, such as the hemispheric 500-mb heights, is a much bigger and more difficult problem. A single atmospheric field might be represented by the values of thousands of 500-mb heights at regularly spaced locations, or gridpoints. Construction of forecasts including probabilities for all these heights and their relationships (e.g., correlations) with heights at the other gridpoints is a very big task, and in practice only approximations to their complete probability description have been achieved. Expressing and communicating aspects of the large amounts of information in a probabilistic field forecast pose further difficulties.

7.6.2. Stochastic Dynamical Systems in Phase Space

Much of the conceptual basis for probabilistic field forecasting is drawn from Gleeson (1961, 1970), who noted analogies to quantum and statistical mechanics; and Epstein (1969c), who presented both theoretical and practical approaches to the problem of uncertainty in (simplified) dynamical weather forecasts. In this approach, which Epstein called *stochastic dynamic prediction*, the physical laws governing the motions and evolution of the atmosphere are regarded as deterministic. However, in practical problems the equations that describe these laws must operate on initial values that are not known with certainty and that therefore can be described by a joint probability distribution. Conventional deterministic forecasts use the dynamical governing equations to describe the future evolution of a single initial state that is regarded as the true initial state. The idea behind stochastic dynamic forecasts is to allow the deterministic governing equations to operate on the probability distribution describing the uncertainty about the initial state of the atmosphere. In principle this process yields, as forecasts, probability distributions describing uncertainty about the future state of the atmosphere. (But actually, since the dynamical models are not perfect representations of the real atmosphere, their imperfections further contribute to forecast uncertainty, as detailed more fully in Section 7.7.)

Visualizing or even conceptualizing the initial and forecast probability distributions is difficult, especially when they involve joint probabilities pertaining to large numbers of forecast variables. This visualization or conceptualization is most commonly and easily done using the concept of a *phase space*. A phase space is a geometrical representation of the hypothetically possible states of a dynamical system, where each of the coordinate axes defining this geometry pertains to one of the forecast variables of the system. Within the phase space, a "state" of the dynamical system is defined by specification of particular values for each of these forecast variables, and therefore corresponds to a single point in this (generally high-dimensional) space.

For example, a simple dynamical system that is commonly encountered in textbooks on physics or differential equations is the swinging pendulum. The state of the dynamics of a pendulum can be completely described by two variables: its angular position and its velocity. At the extremes of the pendulum's arc, its angular position is maximum (positive or negative) and its velocity is zero. At the bottom of its arc, the angular position of the swinging pendulum is zero and its speed (corresponding to either a positive or negative velocity) is maximum. When the pendulum finally stops, both its angular position and velocity are zero. Because the motions of a pendulum can be fully described by two variables, its phase space is two-dimensional. That is, its phase space is a phase-plane. The changes through time of the state of the pendulum system can be described by a path, known as an *orbit*, or a *trajectory*, on this phase-plane.

Figure 7.23 shows the trajectory of a hypothetical pendulum in its phase space. That is, this figure is a graph in phase space of the motions of a pendulum and their changes through time. The trajectory begins at the single point corresponding to the initial state of the pendulum: it is dropped from the right with zero initial velocity (A). As it drops, it accelerates and acquires leftward velocity, which increases until the pendulum passes through the vertical position (B). The pendulum then decelerates, slowing until it stops at its maximum left position (C). As the pendulum drops again it moves to the right, stopping short of its initial position because of friction (D). The pendulum continues to swing back and forth until it finally comes to rest in the vertical position (E).

The phase space of an atmospheric model has many more dimensions than that of the pendulum system. Epstein (1969c) considered a highly simplified model of the atmosphere having only eight variables. Its phase space was therefore eight-dimensional, which is small but still much too big to imagine explicitly. The phase spaces of operational weather forecasting models typically have millions of dimensions, each corresponding to one of the millions of variables [(horizontal gridpoints) x (vertical levels) x (prognostic variables)] represented. The trajectory of the atmosphere or a model of the atmosphere is also qualitatively more complicated than that of the pendulum because it is not attracted to a single point in the phase space, as is the pendulum trajectory in Figure 7.23. Also very importantly, the pendulum dynamics do not exhibit the sensitivity to initial conditions that has come to be known as chaotic behavior, or *chaos*. Releasing the pendulum slightly further to the right or left relative to its initial point in Figure 7.23, or with a slight upward or downward push, would produce a very similar trajectory that would track the spiral in Figure 7.23 very closely and arrive at the same place in the center of the diagram at nearly the same time. The corresponding behavior of the

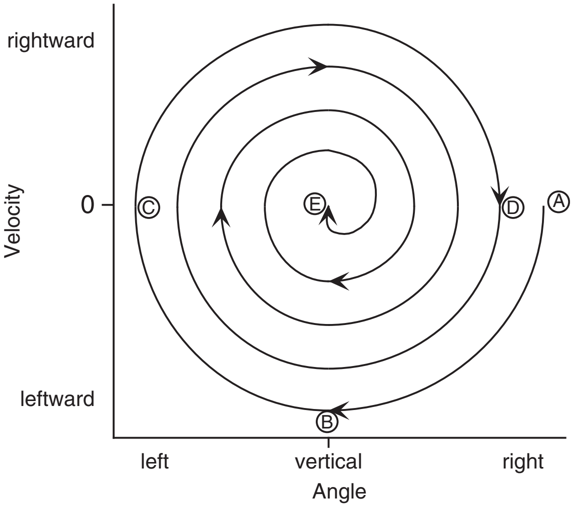


FIGURE 7.23 Trajectory of a swinging pendulum in its two-dimensional phase space, or phase-plane. The pendulum has been dropped from position (A) on the right, from which point it swings in arcs of decreasing angle. Finally, it slows to a stop, with zero velocity in the vertical position (E).

atmosphere, or of a realistic mathematical model of it, would be quite different. Nevertheless, the changes in the flow within a model atmosphere through time can still be imagined abstractly as a trajectory through its multidimensional phase space.

The uncertainty about the initial state of the atmosphere, from which a dynamical model is initialized, can be conceived of as a probability distribution in its phase space. In a two-dimensional phase space like the one shown in Figure 7.23, we might imagine a bivariate normal distribution (Section 4.4.2), with ellipses of constant probability describing the spread of plausible initial states around the best guess, or mean value. Alternatively, we can imagine a cloud of points around the mean value, whose density (number of points per unit area) decreases with distance from the mean. In a three-dimensional phase space, the distribution might be imagined as a cigar- or blimp-shaped cloud of points, again with density decreasing with distance from the mean value. Higher-dimensional spaces cannot be visualized explicitly, but probability distributions within them can be imagined by analogy.

In concept, a stochastic dynamic forecast moves the probability distribution of the initial state through the phase space as the forecast is advanced in time, according to the laws of fluid dynamics represented in the model equations. However, trajectories in the phase space of a dynamical model (or of the real atmosphere) are not nearly as smooth and regular as the pendulum trajectory shown in Figure 7.23. As a consequence, the shape of the initial distribution is stretched and distorted as the forecast is advanced. It will tend to become more dispersed at longer forecast lead times, reflecting the increased uncertainty of forecasts further into the future. Furthermore, these trajectories are not attracted to a single point as are pendulum trajectories in the phase space of Figure 7.23. Rather, the *attractor*, or set of points in the phase space that can be visited after an initial transient period, is a rather complex geometrical object. A single point in the phase space of an atmospheric model corresponds to a unique weather situation, and the collection of these possible points that constitutes the attractor can be interpreted as the climate of the dynamical model. This set of allowable states occupies only a small fraction of the (hyper-) volume of the phase space, since many combinations of atmospheric variables will be physically impossible or dynamically inconsistent.

Equations describing the evolution of the initial-condition probability distribution can be derived through introduction of a continuity, or conservation, equation for probability (Ehrendorfer, 1994, 2006; Gleeson, 1970). However, the dimensionality of phase spaces for problems of practical forecasting interest are too large to allow direct solution of these equations. Epstein (1969c) introduced a simplification that rests on a restrictive assumption about the shapes of the probability distributions in phase space, which is expressed in terms of the moments of their distributions. However, even this approach is impractical for all but the simplest atmospheric models.

7.6.3. Ensemble Forecasts

The practical solution to the analytic intractability of sufficiently detailed stochastic dynamic equations is to approximate these equations using Monte Carlo methods, as proposed by Leith (1974) and now called *ensemble forecasting*. These Monte Carlo solutions bear the same relationship to stochastic dynamic forecast equations as the Monte Carlo resampling tests introduced in Section 5.3.3 bear to the analytical tests they approximate. (Recall that resampling tests are appropriate and useful in situations where the underlying mathematics are difficult or impossible to evaluate analytically.) Lewis (2005) traces the history of this confluence of dynamical and statistical ideas in atmospheric prediction. Reviews of current operational use of the ensemble forecasting approach can be found in Buizza et al. (2005), Cheung (2001), and Kalnay (2003).

The ensemble forecast procedure begins in principle by drawing a finite sample from the probability distribution describing the uncertainty of the initial state of the atmosphere. Imagine that a few members of the point cloud surrounding the mean estimated atmospheric state in phase space are picked randomly. Collectively, these points are called the ensemble of initial conditions, and each represents a plausible initial state of the atmosphere consistent with the uncertainties in observation and analysis. Rather than explicitly predicting the movement of the entire initial-state probability distribution through the phase space of the dynamical model, that movement is approximated by the collective trajectories of the ensemble of sampled initial points. It is for this reason that the Monte Carlo approximation to stochastic dynamic forecasting is known as ensemble forecasting. Each of the points in the initial ensemble provides the initial conditions for a separate dynamical integration. At the initial time, all the ensemble members are very similar to each other. The distribution in phase space of this ensemble of points after the forecasts have been advanced to a future time then approximates how the full true initial probability distribution would have been transformed by the governing physical laws that are expressed in the dynamics of the model.

Figure 7.24 illustrates the nature of ensemble forecasting in an idealized two-dimensional phase space. The circled X in the initial-time ellipse represents the single best initial value, from which a conventional deterministic dynamical integration would begin. Recall that, for a real model of the atmosphere, this initial point defines a full set of meteorological maps for all of the variables being forecast. The evolution of this single forecast in the phase space, through an intermediate forecast lead time and to a final forecast lead time, is represented by the heavy solid lines. However, the position of this point in phase space at the initial time represents only one of the many plausible initial states of the atmosphere consistent with errors in the analysis. Around it are other plausible states, which sample the probability distribution for states of the atmosphere at the initial time. This distribution is represented by the small ellipse. The open circles in this ellipse represent eight other members of this

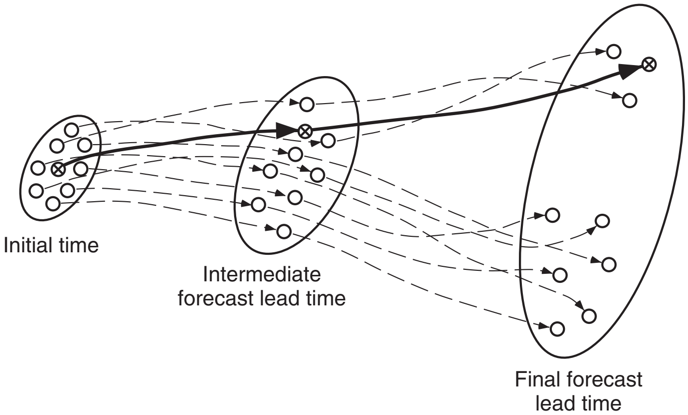


FIGURE 7.24 Schematic illustration of some concepts in ensemble forecasting, plotted in terms of an idealized two-dimensional phase space. The heavy line represents the evolution of the single best analysis of the initial state of the atmosphere, corresponding to the more traditional single deterministic forecast. The dashed lines represent the evolution of individual ensemble members. The ellipse in which they originate represents the probability distribution of initial atmospheric states, which are very close to each other. At the intermediate lead time, all the ensemble members are still reasonably similar. By the final lead time, some of the ensemble members have undergone a regime change and represent qualitatively different flows. Any of the ensemble members, including the solid line, are plausible trajectories for the evolution of the real atmosphere, and there is no way of knowing in advance which will represent the real atmosphere most closely.

distribution. This ensemble of nine initial states approximates the variations represented by the full distribution from which they were drawn.

The Monte Carlo approximation to a stochastic dynamic forecast is constructed by repeatedly running the dynamical model, once for each of the members of the initial ensemble. The trajectories through the phase space of each of the ensemble members are only modestly different at first, indicating that all nine integrations represented in Figure 7.24 are producing fairly similar forecasts at the intermediate lead time. Accordingly, the probability distribution describing uncertainty about the state of the atmosphere at the intermediate lead time would not be a great deal larger than at the initial time. However, between the intermediate and final lead times the trajectories diverge markedly, with three (including the one started from the central value of the initial distribution) producing forecasts that are similar to each other and the remaining six members of the ensemble predicting rather different atmospheric states at that time. The underlying distribution of uncertainty that was fairly small at the initial time has been stretched substantially, as represented by the large ellipse at the final lead time. The dispersion of the ensemble members at that time allows the nature of that distribution to be estimated and is indicative of the uncertainty of the forecast, assuming that the dynamical model includes only negligible errors in the representations of the governing physical processes. If only the single forecast started from the best initial condition had been made, this information would not be available.

7.6.4. Choosing Initial Ensemble Members

Ideally, we would like to produce ensemble forecasts based on a large number of possible initial atmospheric states drawn randomly from the PDF of initial-condition uncertainty in phase space. However, each member of an ensemble of forecasts is produced by a complete rerunning of the dynamical model, each of which requires a substantial amount of computing. As a practical matter, computer time is a limiting factor at operational forecast centers, and each center must make a subjective judgment balancing the number of ensemble members to include in relation to the spatial resolution of the model used to integrate them forward in time. Consequently, the sizes of operational forecast ensembles are limited, and it is important that initial ensemble members be chosen well. Their selection is further complicated by the fact that the initial-condition PDF in phase space is unknown, and it presumably changes from day to day, so that the ideal of simple random samples from this distribution cannot be achieved in practice.

The simplest, and historically first, method of generating initial ensemble members was to begin with a best analysis, assumed to be the mean of the probability distribution representing the uncertainty of the initial state of the atmosphere. Variations around this mean state can be easily generated by adding random numbers characteristic of the errors or uncertainty in the instrumental observations underlying the analysis (Leith, 1974). For example, these random values might be Gaussian variates with zero mean, implying an unbiased combination of measurement and analysis errors. In practice, however, simply adding independent random numbers to a single initial field has been found to yield ensembles whose members are too similar to each other, probably because much of the variation introduced in this way is dynamically inconsistent, so that the corresponding energy is quickly dissipated in the model (Palmer et al., 1990). The consequence is that the dispersion of the resulting forecast ensemble underestimates the uncertainty in the forecast.

As of the time of this writing (2010), there are three dominant methods of choosing initial ensemble members in operational practice. In the United States, the National Centers for Environmental Prediction use the *breeding method* (Ehrendorfer, 1997; Kalnay, 2003; Toth and Kalnay, 1993, 1997). In this approach, differences in the three-dimensional patterns of the predicted variables, between the ensemble members and the single "best" (control) analysis, are chosen to look like differences between recent forecast ensemble members and the forecast from the corresponding previous control analysis. The patterns are then scaled to have magnitudes appropriate to analysis uncertainties. These bred patterns are different from day to day and emphasize features with respect to which the ensemble members are diverging most rapidly. The breeding method is relatively inexpensive computationally.

In contrast, the European Centre for Medium-Range Weather Forecasts generates initial ensemble members using *singular vectors* (Buizza, 1997; Ehrendorfer, 1997; Kalnay, 2003; Molteni et al., 1996). Here the fastest growing characteristic patterns of differences from the control analysis in a linearized version of the full forecast model are calculated, again for the specific weather situation of a given day. Linear combinations (in effect, weighted averages) of these patterns, with magnitudes reflecting an appropriate level of analysis uncertainty, are then added to the control analysis to define the ensemble members. Ehrendorfer and Tribbia (1997) present theoretical support for the use of singular vectors to choose initial ensemble members, although its use requires substantially more computation than does the breeding method.

The Meteorological Service of Canada generates its initial ensemble members using a method called the *ensemble Kalman filter (*EnKF*)* (Houtekamer and Mitchell, 2005). This method is related to the multivariate extension of conjugate Bayesian updating of a Gaussian prior distribution (Section 6.3.4). Here the ensemble members from the previous forecast cycle define the Gaussian prior distribution, and the ensemble members are updated using a Gaussian likelihood function (i.e., data-generating process) for available observed data assuming known data variance (characteristic of the measurement errors), to yield new initial ensemble members from a Gaussian posterior distribution. The initial ensembles are relatively compact as a consequence of their (posterior) distribution being constrained by the observations, but the ensemble members diverge as each is integrated forward in time by the dynamical model, producing a more dispersed prior distribution for the next update cycle. Expositions and literature reviews for the EnKF are provided by Evensen (2003) and Hamill (2006).

In the absence of direct knowledge about the PDF of initial-condition uncertainty, how best to define initial ensemble members is not completely clear and is the subject of ongoing research. Comparisons of the methods just described using simplified, idealized dynamical models (Bowler, 2006a; Descamps and Talagrand, 2007) have indicated better results with the EnKF. However, to date the methods do not appear to have been compared in a full dynamical model of operational complexity.

7.6.5. Ensemble Average and Ensemble Dispersion

One simple application of ensemble forecasting is to average the members of the ensemble in order to obtain a single forecast. The motivation is to obtain a forecast that is more accurate than the single forecast initialized with the best estimate of the initial state of the atmosphere. Epstein (1969a) pointed out that the time-dependent behavior of the ensemble mean is different from the solution of forecast equations using the initial mean value, and concluded that in general the best forecast is not the single forecast initialized with the best estimate of initial conditions. The first of these conclusions, at least, should not be surprising since a dynamical model is in effect a highly nonlinear function that transforms a set of initial atmospheric conditions to a set of forecast atmospheric conditions.

In general, the average of a nonlinear function over some set of particular values of its argument is not the same as the function evaluated at the average of those values. That is, if the function $f(x)$ is nonlinear,

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) \neq f\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right). \tag{7.41}$$

To illustrate simply, consider the three values $x_1 = 1$, $x_2 = 2$, and $x_3 = 3$. For the nonlinear function $f(x) = x^2 + 1$, the left side of Equation 7.41 is 5 2/3, and the right side of that equation is 5. We can easily verify that the inequality of Equation 7.41 holds for other nonlinear functions (e.g., $f(x) = \log(x)$ or $f(x) = 1/x$) as well. By contrast, for the linear function $f(x) = 2x + 1$ the two sides of Equation 7.41 are both equal to 5.

Extending this idea to ensemble forecasting, we might like to know the atmospheric state corresponding to the center of the ensemble in phase space for some time in the future. This central value of the ensemble will approximate the center of the stochastic dynamic probability distribution at that future time, after the initial distribution has been transformed by the nonlinear forecast equations. The Monte Carlo approximation to this future value is the ensemble average forecast. The ensemble average forecast is obtained simply by averaging together the ensemble members for the lead time of interest, which corresponds to the left side of Equation 7.41. By contrast, the right side of Equation 7.41 represents the single forecast started from the average initial value of the ensemble members. Depending on the nature of the initial distribution and on the dynamics of the model, this single forecast may or may not be close to the ensemble average forecast.

In the context of weather forecasts, the benefits of ensemble averaging appear to derive primarily from averaging out elements of disagreement among the ensemble members, while emphasizing features that generally are shared by the members of the forecast ensemble. Particularly for longer lead times, ensemble average maps tend to be smoother than instantaneous snapshots and so may seem unmeteorological, or more similar to smooth climatic averages. Palmer (1993) suggests that ensemble averaging will improve the forecast only until a regime change, or a change in the long-wave pattern, takes place, and he illustrates this concept nicely using the simple Lorenz (1963) model. This problem also is illustrated in Figure 7.24, where a regime change is represented by the bifurcation of the trajectories of the ensemble members between the intermediate and final lead times. At the intermediate lead time, before some of the ensemble members undergo this regime change, the center of the distribution of ensemble members is well represented by the ensemble average, which is a better central value than the single member of the ensemble started from the "best" initial condition. At the final forecast lead time the distribution of states has been distorted into two distinct groups. Here the ensemble average will be located somewhere in the middle, but near none of the ensemble members.

A particularly important aspect of ensemble forecasting is its capacity to yield information about the magnitude and nature of the uncertainty in a forecast. In principle the forecast uncertainty is different on different forecast occasions, and this notion can be thought of as state-dependent predictability. The value to forecast users of communicating the different levels of forecast confidence that exist on different occasions was recognized early in the twentieth century (Cooke, 1906b; Murphy, 1998). Qualitatively, we have more confidence that the ensemble mean is close to the eventual state of the atmosphere if the dispersion of the ensemble is small. Conversely, if the ensemble members are all very different from each other, the future state of the atmosphere is more uncertain. One approach to "forecasting forecast skill" (Ehrendorfer, 1997; Kalnay and Dalcher, 1987; Palmer and Tibaldi, 1988) is to anticipate the accuracy of a forecast as being inversely related to the dispersion of the ensemble members. Operationally, forecasters do this informally when comparing the results from different dynamical models, or when comparing successive forecasts for a particular time in the future that were initialized on different days.

More formally, the *spread-skill relationship* for a collection of ensemble forecasts often is characterized by the correlation, over a collection of forecast occasions, between some measure of the ensemble spread such as the variance or standard deviation of the ensemble members around their ensemble mean on each occasion, and a measure of the predictive accuracy of the ensemble mean on that occasion. The accuracy is often characterized using either the mean squared error (Equation 8.30) or its square root, although other measures have been used in some studies. These spread-skill correlations generally have been found to be fairly modest and rarely exceed 0.5, which corresponds to accounting for 25% or less of the accuracy variations (e.g., Atger, 1999; Grimit and Mass, 2002; Hamill et al. 2004; Whittaker and Loughe, 1998), although some more recently reported values (e.g., Sherrer et al., 2004; Stensrud and Yussouf, 2003) have been higher. Figure 7.25 shows forecast accuracy, as measured by average root-mean squared error (RMSE) of ensemble members, as functions of ensemble spread measured by average root-mean squared differences among all possible pairs of the ensemble members, for forecasts of 500-mb height over western Europe by the 51-member ECMWF ensemble prediction system for June 1997–December 2000. Clearly the more accurate forecasts (smaller RMSE) tend to be associated with smaller ensemble spreads, and vice versa, with this relationship being stronger for the shorter, 96-hour lead time.

Alternative approaches to characterizing the spread-skill relationship continue to be investigated. Moore and Kleeman (1998) calculate probability distributions for forecast skill, conditional on ensemble spread. Toth et al. (2001) present an interesting alternative characterization of the ensemble dispersion, in terms of counts of ensemble forecasts between climatological deciles for the predictand. Tang et al. (2008) consider predicting forecast skill using information-theoretic characterizations of

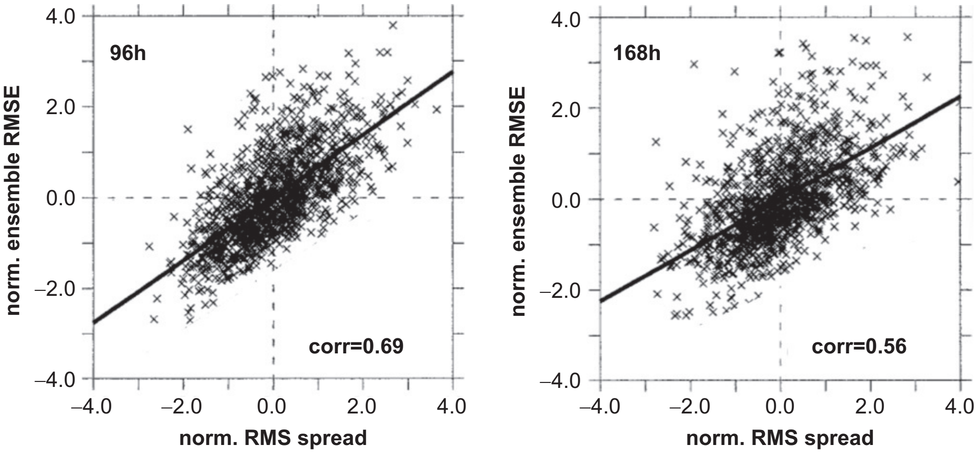


FIGURE 7.25 Scatterplots and correlations between forecast accuracy (vertical) and ensemble spread (horizontal) for ECMWF 500-mb height forecasts over western Europe, 1997–2000, at 96-h and 168-h lead times. *Modified from Sherrer* et al. *(2004).*

the forecast ensemble. Some other promising alternative characterizations of the ensemble spread have been proposed by Ziehmann (2001).

7.6.6. Graphical Display of Ensemble Forecast Information

A prominent attribute of ensemble forecast systems is that they generate large amounts of multivariate information. As noted in Section 3.6, the difficulty of gaining even an initial understanding of a new multivariate data set can be reduced through the use of well-designed graphical displays. It was recognized early in the development of what is now ensemble forecasting that graphical display would be an important means of conveying the resulting complex information to forecasters (Epstein and Fleming, 1971; Gleeson, 1967), and operational experience is still accumulating regarding the most effective means of doing so. This section summarizes current practice according to three general types of graphics: displays of raw ensemble output or selected elements of the raw output; displays of statistics summarizing the ensemble distribution; and displays of ensemble relative frequencies for selected predictands. Displays based on more sophisticated statistical analysis of an ensemble are also possible (e.g., Stephenson and Doblas-Reyes, 2000).

Perhaps the most direct way to visualize an ensemble of forecasts is to plot them simultaneously. Of course, for even modestly sized ensembles each element (corresponding to one ensemble member) of such a plot must be small in order for all the ensemble members to be viewed simultaneously. Such collections are called *stamp maps* because each of its individual component maps is sized approximately like a postage stamp, allowing only the broadest features to be discerned. For example, Figure 7.26 shows 51 stamp maps from the ECMWF ensemble prediction system, for surface pressure over western Europe ahead of a large and destructive winter storm that occurred in December 1999. The ensemble consists of 50 members, plus the control forecast begun at the “best” initial atmospheric state, labeled “Deterministic predictions.” The subsequently analyzed surface pressure field, labeled “Verification,” indicates a deep, intense surface low centered near Paris. The control forecast missed

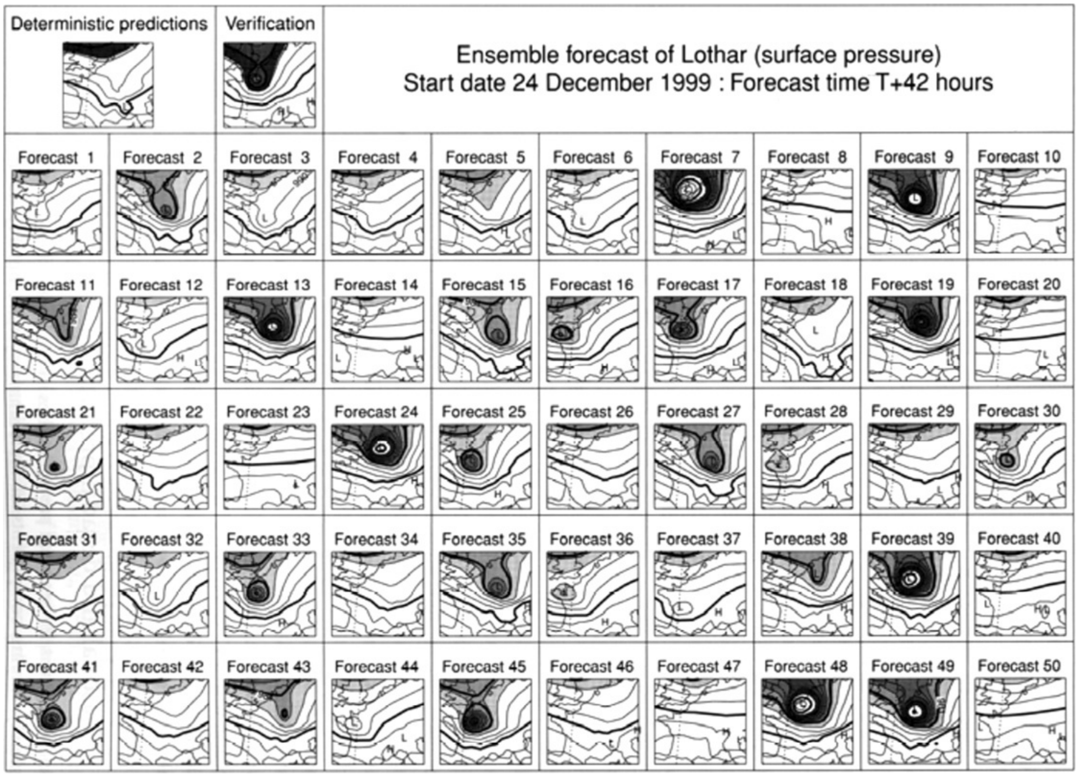


FIGURE 7.26 Stamp maps from the 51-member ECMWF ensemble forecast for surface pressure over western Europe. The Verification shows the corresponding surface analysis 42 h later during winter storm Lothar. *From Palmer* et al. *(2001).*

this important feature completely, as did many of the ensemble members. However, a substantial number of the ensemble members did portray a deep surface low, suggesting a substantial probability for this destructive storm, 42 h in advance. Although fine details of the forecast are difficult, if not impossible, to discern from the small images in a stamp map, a forecaster with experience in the interpretation of this kind of display can get an overall sense of the outcomes that are plausible, according to this sample of ensemble members. A further step that sometimes is taken with a collection of stamp maps is to group them objectively into subsets of similar maps using a cluster analysis (see Section 15.2).

Part of the difficulty in interpreting a collection of stamp maps is that the many individual displays are difficult to comprehend simultaneously. Superposition of a set of stamp maps would alleviate this difficulty if not for the problem that the resulting plot would be too cluttered to be useful. However, seeing each contour of each map is not necessary to form a general impression of the flow. Indeed, seeing only one or two well-chosen pressure or height contours is often sufficient to define the main features, since typically the contours roughly parallel each other. Superposition of one or two well-selected contours from each of the stamp maps often does yield a sufficiently uncluttered composite to be interpretable, which is known as the *spaghetti plot*. Figure 7.27 shows three spaghetti plots for the 5520-m contour of the 500-mb surface over North America, as forecast 12, 36, and 84 hours

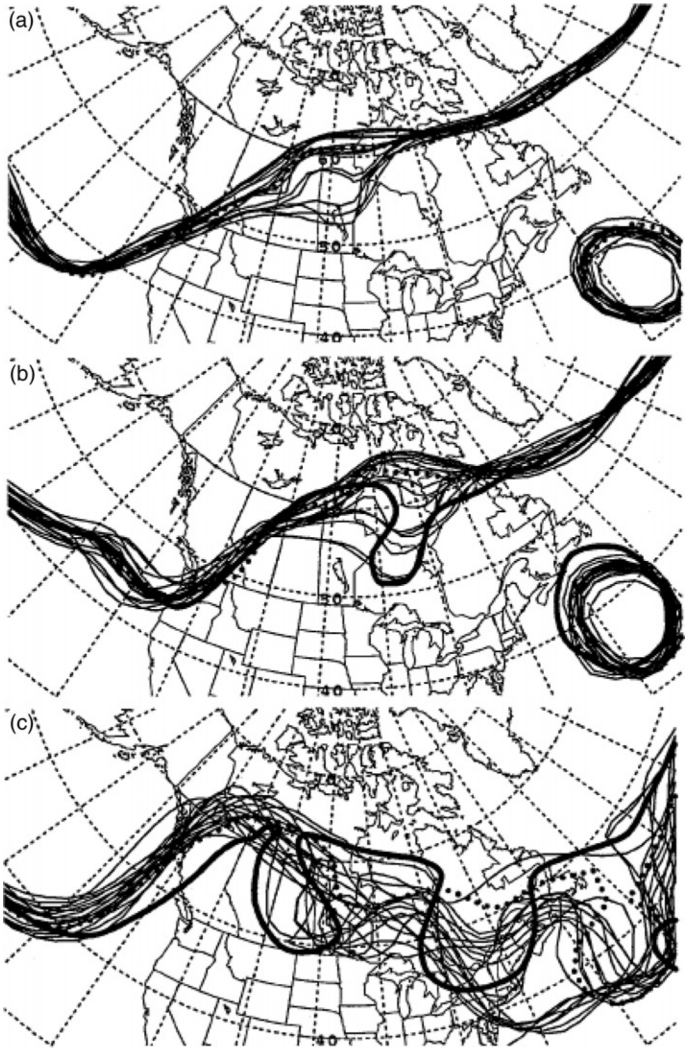


FIGURE 7.27 Spaghetti plots for the 5520-m contour of the 500-mb height field over North America forecast by the National Centers for Environmental Prediction, showing forecasts for (a) 12 h, (b) 36 h, and (c) 84h after the initial time of 0000 UTC, March 14, 1995. Light lines show the contours produced by each of the 17 ensemble members, the dotted line shows the control forecast, and the heavy lines in panels (b) and (c) indicate the verifying analyses. *From Toth* et al.*, 1997.*

after the initial time of 0000 UTC, March 14, 1995. In Figure 7.27a the 17 ensemble members generally agree quite closely for the 12-hour forecast, and even with only the 5520-m contour shown the general nature of the flow is clear: the trough over the eastern Pacific and the cutoff low over the Atlantic are clearly indicated.

At the 36-hour lead time (Figure 7.27b) the ensemble members are still generally in close agreement about the forecast flow, except over central Canada, where some ensemble members produce

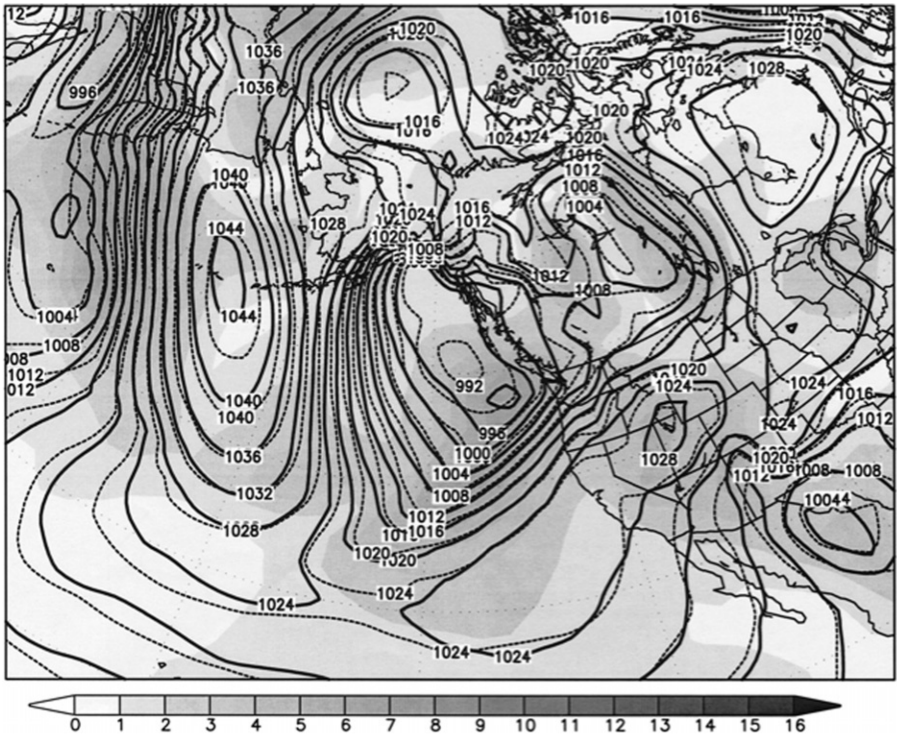


FIGURE 7.28 Ensemble mean (solid) and ensemble standard deviation (shading) for a 12-hour forecast of sea-level pressure, valid 0000 UTC, January 29, 1999. Dashed contours indicate the single control forecast. *From Toth* et al. *(2001).*

a short-wave trough. The 500-mb field over most of the domain would be regarded as fairly certain except in this area, where the typical interpretation would be a substantial but not dominant probability of a short-wave feature that was missed by the single forecast from the control analysis (dotted). The heavy line in this panel indicates the subsequent analysis for the 36-hour lead time. At the 84-hour lead time (Figure 7.27c) there is still substantial agreement about (and thus relatively high probability would be inferred for) the trough over the Eastern Pacific, but the forecasts for the continent and the Atlantic have begun to diverge quite strongly, suggesting the pasta dish for which this kind of plot is named. Spaghetti plots have proven to be quite useful in visualizing the evolution of the forecast flow, simultaneously with the dispersion of the ensemble. The effect is even more striking when a series of spaghetti plots is animated, which can be appreciated at some operational forecast center websites.

It can be informative to condense the large amount of information from an ensemble forecast into a small number of summary statistics and to plot maps of these. By far the most common such plot, suggested initially by Epstein and Fleming (1971), is simultaneous display of the ensemble mean and standard deviation fields. That is, at each of a number of gridpoints the average of the ensemble members is calculated, as well as the standard deviation of the ensemble members around this average. Figure 7.28 is one such plot, for a 12-hour forecast of sea-level pressure (mb) over much of North America and the north Pacific, valid at 0000 UTC, January 29, 1999. Here the solid contours represent the ensemble mean field, and the shading indicates the field of ensemble standard deviations.

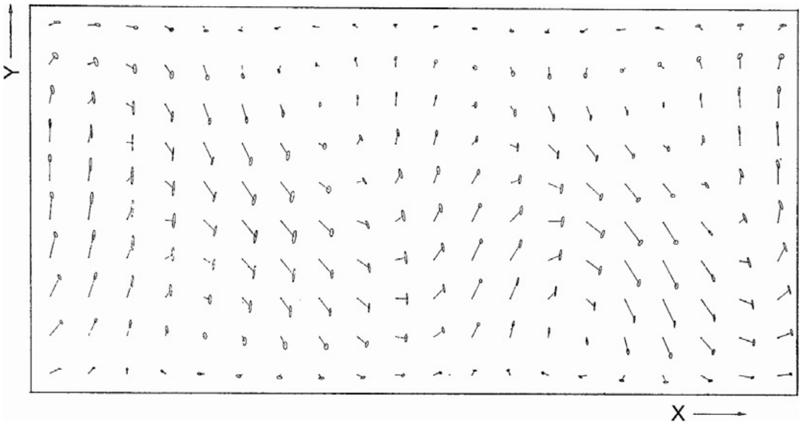


FIGURE 7.29 A forecast wind field from an idealized modeling experiment, expressed in probabilistic form. Line lengths and orientations show forecast mean wind vectors directed from the gridpoint locations to the ellipses. The ellipses indicate boundaries containing the observed wind with probability 0.50. *From Epstein and Fleming (1971).*

These standard deviations indicate that the anticyclone over eastern Canada is predicted quite consistently among the ensemble members (ensemble standard deviations generally less than 1 mb), and the pressures in the eastern north Pacific and east of Kamchatka, where large gradients are forecast, are somewhat less certain (ensemble standard deviations greater than 3 mb).

Gleeson (1967) suggested combining maps of forecast u and v wind components with maps of probabilities that the forecasts will be within 10 knots of the eventual observed values. Epstein and Fleming (1971) suggested that a probabilistic depiction of a horizontal wind field could take the form of Figure 7.29. Here the lengths and orientations of the lines indicate the mean of the forecast distributions of wind vectors, blowing from the gridpoints the ellipses. The probability is 0.50 that the true wind vectors will terminate within the corresponding ellipse. It has been assumed in this figure that the uncertainty in the wind forecasts is described by the bivariate normal distribution, and the ellipses have been drawn as explained in Example 11.1. The tendency for the ellipses to be oriented in a north-south direction indicates that the uncertainties of the meridional winds are greater than those for the zonal winds, and the tendency for the larger velocities to be associated with larger ellipses indicates that these wind values are more uncertain.

Ensemble forecasts for surface weather elements at a single location can be concisely summarized by time series of boxplots for selected predictands, in a plot called an *ensemble meteogram*. Each of these boxplots displays the dispersion of the ensemble for one predictand at a particular forecast lead time, and jointly they show the time evolutions of the forecast central tendencies and uncertainties, through the forecast period. Figure 7.30 shows an example from the Japan Meteorological Agency, in which boxplots representing ensemble dispersion for four weather elements at Tsukuba are plotted at 6-hourly intervals. The plot indicates greater uncertainty in the cloud cover and precipitation forecasts, and the increasing uncertainty with increasing lead time is especially evident for the temperature forecasts.

Figure 7.31 shows an alternative to boxplots for portraying the time evolution of the ensemble distribution for a predictand. In this *plume graph* the contours indicate heights of the ensemble dispersion, expressed as a PDF, as a function of time for forecast 500-mb heights over southeast England.

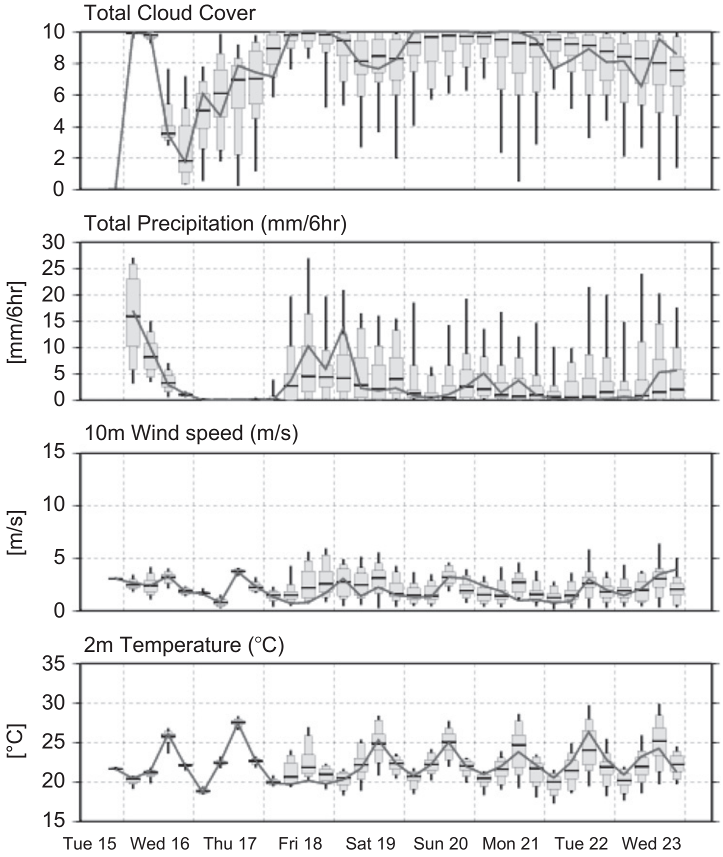


FIGURE 7.30 Ensemble meteogram for Tsukuba, Japan, from a Japan Meteorological Agency forecast ensemble begun on 1200 UTC, June 15, 2010. Wider portions of boxplots indicate the interquartile ranges, narrower box portions show middle 80% of the ensemble distributions, and whiskers extend to most extreme ensemble members. Solid line shows the control forecast. *From gpvjma.ccs.hpcc.jp*.

The ensemble can be seen to be quite compact early in the forecast and expresses a large degree of uncertainty by the end of the period.

Finally, information from ensemble forecasts is very commonly displayed as maps of ensemble relative frequencies for dichotomous events, which are often defined according to a threshold for a continuous variable. Ideally, ensemble relative frequency would correspond closely to forecast probability; but because of nonideal sampling of initial ensemble members, together with inevitable deficiencies in the dynamical models used to integrate them forward in time, this interpretation is not literally warranted (Allen et al., 2006; Hansen, 2002; Smith, 2001), and such probability estimates can be improved by applying MOS methods to ensemble forecasts (Section 7.7).

Figure 7.32 shows an example of a very common plot of this kind, for ensemble relative frequency of more than 2 mm of precipitation over 12 hours, at lead times of (a) 7 days, (b) 5 days, and (c) 3 days ahead of the observed event (d). As the lead time decreases, the areas with appreciable forecast probability become more compactly defined and exhibit the generally larger relative frequencies

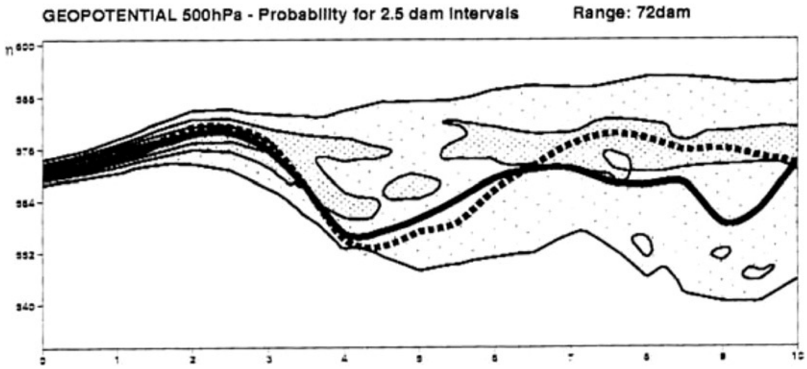


FIGURE 7.31 A plume graph, indicating probability density as a function of time, for a 10-day forecast of 500-mb height over southeast England, initiated 1200 UTC, August 26, 1999. The dashed line shows the high-resolution control forecast, and the solid line indicates the lower-resolution ensemble member begun from the same initial condition. *From Young and Carroll (2002).*

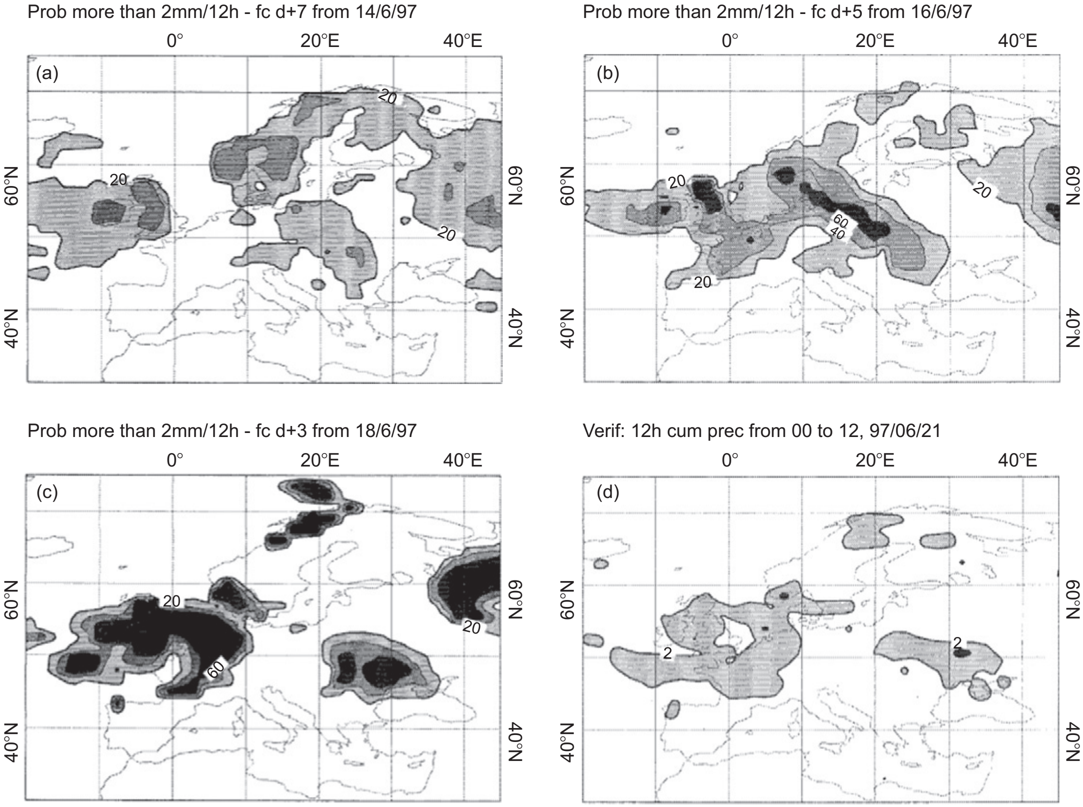

Prob more than 2mm/12h - fc d+7 from 14/6/97
(a)
Prob more than 2mm/12h - fc d+5 from 16/6/97
(b)
Prob more than 2mm/12h - fc d+3 from 18/6/97
(c)
Verif: 12h cum prec from 00 to 12, 97/06/21
(d)
0°
20°E
40°E
60°N
40°N
20
40
60
2

FIGURE 7.32 Ensemble relative frequency for accumulation of >2 mm precipitation over Europe in a 12-hour period (a) 7 days, (b) 5 days, and (c) 3 days ahead of (d) the observed events, on June 21, 1997. Contour interval in (a)–(c) is 0.2. *From Buizza* et al. *(1999a).*

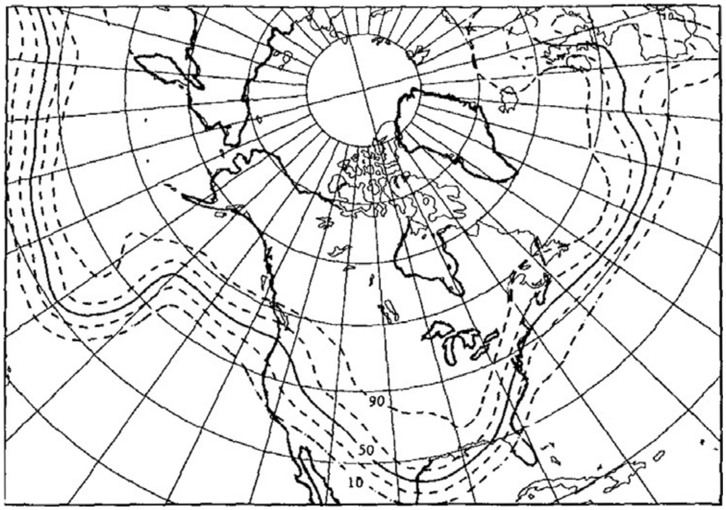


FIGURE 7.33 Ensemble relative frequencies of 1000–500 mb thicknesses less than 5400 m for March 14, 1993 over North America, as estimated using 14 ensemble forecast members. *From Tracton and Kalnay (1993).*

indicative of greater confidence in the event outcomes. Other kinds of probabilistic field maps are also possible, many of which may be suggested by the needs of particular forecast applications. Figure 7.33, showing relative frequencies of forecast 1000–500 mb thickness over North America being less than 5400 m, is one such possibility. Forecasters often use this thickness value as an expected dividing line between rain and snow. At each gridpoint, the fraction of ensemble members predicting 5400-m thickness or less has been tabulated and plotted. Clearly, similar maps for other thickness values could be constructed as easily. Figure 7.33 indicates a relatively high confidence that the cold-air outbreak over the eastern United States will bring air sufficiently cold to produce snow as far south as the Gulf coast.

7.6.7. Effects of Model Errors

Given a perfect dynamical model, integrating a random sample from the PDF of initial-condition uncertainty forward in time would yield a sample from the PDF characterizing forecast uncertainty. Of course, dynamical models are not perfect, so that even if an initial-condition PDF could be known and correctly sampled from, the distribution of a forecast ensemble can at best be only an approximation to a sample from the true PDF for the forecast uncertainty (Hansen, 2002; Palmer, 2006; Smith, 2001).

Leith (1974) distinguished two kinds of model errors. The first derives from the models inevitably operating at a lower resolution than the real atmosphere or, equivalently, occupying a phase space of much lower dimension (Judd et al., 2008). Although still significant, this problem has been gradually addressed and partially ameliorated over the history of dynamical forecasting through progressive increases in model resolution. The second kind of model error derives from the fact that certain physical processes—prominently those operating at scales smaller than the model resolution—are represented incorrectly. In particular, such physical processes (known colloquially in this context as “physics”) generally are represented using some relatively simple function of the explicitly resolved

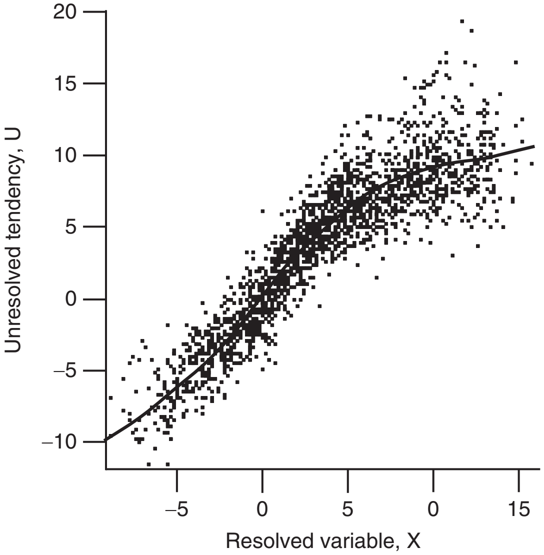


FIGURE 7.34 Scatterplot of the unresolved time tendency, U, of a resolved variable, X, as a function of the resolved variable; together with a regression function representing the conditional average dependence of the tendency on the resolved variable. *From Wilks (2005).*

variables, known as a *parameterization*. Figure 7.34 shows a parameterization (solid curve) for the unresolved part of the tendency (dX/dt) of a resolved variable X, as a function of X itself, in the highly idealized Lorenz '96 (Lorenz, 2006) model (Wilks, 2005). The individual points in Figure 7.34 are a sample of the actual unresolved tendencies, which are summarized by the regression function. In a realistic dynamical model, there are a large number of such parameterizations for various unresolved physical processes, and the effects of these processes on the resolved variables are included in the model as functions of the resolved variables through these parameterizations. It is evident from Figure 7.34 that the parameterization (smooth curve) does not fully capture the range of behaviors for the parameterized process that are actually possible (scatter of points around the curve). Even if the large-scale dynamics have been modeled correctly, nature does not supply the value of the unresolved tendency given by "the" parameterized curve, but rather provides an effectively random realization from the point cloud around it. One way of looking at this kind of model error is that the parameterized physics are not fully determined by the resolved variables. That is, they are uncertain.

One way of representing the errors, or uncertainties, in the parameterized model physics is to extend the idea of the ensemble to include simultaneously a collection of different initial conditions *and* multiple dynamical models (each of which has a different collection of parameterizations). Harrison et al. (1999) found that forecasts using all four possible combinations of two sets of initial conditions and two dynamical model formulations differed significantly, with members of each of the four ensembles clustering relatively closely together, and distinctly from the other three, in the phase space. Other studies (e.g., Hansen, 2002; Houtekamer et al., 1996; Mullen et al., 1999; Mylne et al., 2002a; Stensrud et al., 2000) have found that using such *multimodel ensembles* improves the resulting ensemble forecasts. The components of the Canadian Meteorological Center's operational multimodel ensemble share the same large-scale dynamical formulation, but differ with respect to the structure of various parameterizations (Houtekamer et al., 2009), in effect using different but similar parameterization curves of the kind represented in Figure 7.34, for different ensemble members. A substantial part of the resulting improvement in ensemble performance derives from the multimodel ensembles exhibiting larger ensemble dispersion, so that the ensemble members are less like each other than if an identical dynamical model is used for all forecast integrations. Typically, the dispersion of forecast ensembles is too small (e.g., Buizza, 1997; Stensrud et al., 1999; Toth and Kalnay, 1997), and so expresses too little uncertainty about forecast outcomes (see Section 8.7).

Another approach to capturing uncertainties in the structure of dynamical models is suggested by the scatter around the regression curve in Figure 7.34. From the perspective of Section 7.2, the regression residuals that are differences between the actual (points) and parameterized (regression curve) behavior of the modeled system are random variables. Accordingly, the effects of parameterized processes can be more fully represented in a dynamical model if random numbers are added to the deterministic parameterization function, making the dynamical model explicitly stochastic (e.g., Palmer, 2001; Palmer et al., 2005; Teixeira and Reynolds, 2008). Even if the system being modeled truly does not contain random components, adopting the stochastic view of unresolved, parameterized processes in a dynamical model may improve the resulting forecasts (Judd et al., 2007; Wilks, 2005).

The idea of stochastic parameterizations in dynamical models is not new, having been proposed as early as the 1970s (Lorenz, 1975; Moritz and Sutera, 1981; Pitcher, 1977). However, its use in realistic atmospheric models has been relatively recent (Bowler et al., 2008; Buizza et al., 1999b; Garratt et al., 1990; Lin and Neelin, 2000, 2002; Williams et al., 2003). Particularly noteworthy is the first operational use of a stochastic representation of the effects of unresolved processes in the forecast model at the European Centre for Medium-Range Forecasts, which they called *stochastic physics* and which results in improved forecasts relative to the conventional deterministic parameterizations (Buizza et al., 1999b; Mullen and Buizza, 2001). Stochastic parameterization is still at an early stage of development, and is the subject of ongoing research (e.g., Berner et al., 2010, Neelin et al., 2010, Plant and Craig, 2007, Tompkins and Berner, 2008).

Stochastic parameterizations also have been used in simplified climate models, to represent atmospheric variations on the timescale of weather, beginning the 1970s (e.g., Hasselmann, 1976; Lemke, 1977; Sutera, 1981), and in continuing work (Imkeller and Monahan, 2002; Imkeller and von Storch, 2001). Some relatively recent papers applying this idea to prediction of the El Niño phenomenon are Penland and Sardeshmukh (1995), Saravanan and McWilliams (1998), and Thompson and Battisti (2001).

7.7. ENSEMBLE MOS

7.7.1. Why Ensembles Need Postprocessing

In principle, initial ensemble members chosen at random from the PDF characterizing initial-condition uncertainty, and integrated forward in time with a perfect dynamical model, will produce an ensemble of future atmospheric states that is a random sample from the PDF characterizing forecast uncertainty. Ideally, then, the dispersion of a forecast ensemble characterizes the uncertainty in the forecast, so that small ensemble dispersion (all ensemble members similar to each other) indicates low uncertainty, and large ensemble dispersion (large differences among ensemble members) signals large forecast uncertainty.

In practice, the initial ensemble members are chosen in ways that do not randomly sample from the PDF of initial-condition uncertainty (Section 7.6.4), and errors in the dynamical models deriving mainly from unresolved scales and processes produce errors in ensemble forecasts just as they do in conventional single-integration forecasts. Accordingly, the dispersion of a forecast ensemble can at best only approximate the PDF of forecast uncertainty (Hansen, 2002; Smith, 2001). In particular, a forecast ensemble may reflect errors both in statistical location (most or all ensemble members being well away from the actual state of the atmosphere, but relatively nearer to each other) and dispersion (either under- or overrepresenting the forecast uncertainty). Often, operational ensemble forecasts are found to exhibit too little dispersion (e.g. Buizza, 1997; Buizza et al., 2005; Hamill, 2001; Toth et al., 2001; Wang and Bishop, 2005), which leads to overconfidence in probability assessment if ensemble relative frequencies are interpreted directly as estimating probabilities.

To the extent that ensemble forecast errors have consistent characteristics, they can be corrected through *ensemble MOS* methods that summarize a historical database of these forecast errors, just as is done for single-integration dynamical forecasts. From the outset of ensemble forecasting (Leith, 1974), it was anticipated that use of finite ensembles would yield errors in the forecast ensemble mean that could be statistically corrected using a database of previous errors. MOS postprocessing is a more difficult problem for ensemble forecasts than for ordinary single-integration dynamical forecasts, or for the ensemble mean, because ensemble forecasts are equally susceptible to the ordinary biases introduced by errors and inaccuracies in the dynamical model formulation, in addition to their usual underdispersion bias. Either or both of these kinds of problems in ensemble forecasts can be corrected using MOS methods.

Ultimately the goal of ensemble MOS methods is to estimate a forecast PDF or CDF on the basis of the discrete approximation provided by a finite, n_{ens}-member ensemble. If the effects of initial-condition and model errors were not important, this task could be accomplished by operating only on the ensemble members at hand, without regard to the statistical characteristics of past forecast errors. Probably the simplest such non-MOS approach is to regard the forecast ensemble as a random sample from the true forecast CDF, and estimate cumulative probabilities from that CDF using a plotting position estimator (Section 3.3.7). The most commonly used, though usually suboptimal, such estimator is the *democratic voting* method. Denoting the quantity being forecast, or verification, as V, and the distribution quantile whose cumulative probability is being estimated as q, this method computes

$$\Pr\{V \le q\} = \frac{1}{n_{ens}} \sum_{i=1}^{n_{ens}} I(x_i \le q) = \frac{rank(q) - 1}{n_{ens}}, \tag{7.42}$$

where the indicator function $I(\bullet) = 1$ if its argument is true and is zero otherwise, and rank(q) indicates the rank of the quantile of interest in a hypothetical $n_{ens} + 1$ member ensemble consisting of the ensemble members x_i and that quantile. Equation 7.42 is equivalent to the Gumbel plotting position estimator (Table 3.2) and has the unfortunate property of assigning zero probability to any quantile less than the smallest ensemble member, $x_{(1)}$, and unit probability to any quantile greater than the largest ensemble member, $x_{(n_{ens})}$. Other plotting position estimators do not have these deficiencies; for example, using the Tukey plotting position (Wilks, 2006b),

$$\Pr\{V \le q\} = \frac{Rank(q) - 1/3}{(n_{ens} + 1) + 1/3}. \tag{7.43}$$

Katz and Ehrendorfer (2006) derive a cumulative probability estimator equivalent to the Weibull plotting position using a conjugate Bayesian analysis (Section 6.3.2) with a uniform prior distribution and a binomial likelihood for the ensemble members' binary forecasts of q. However, the cumulative probability estimators in Equation 7.42 and 7.43 will still lead to inaccurate, overconfident results unless the ensemble size is large or the forecasts are reasonably skillful, even if the ensemble is free of bias errors and exhibits dispersion that is consistent with the actual forecast uncertainty (Richardson, 2001; see Section 8.7).

7.7.2. Regression Methods

Direct transformation of a collection of ensemble forecasts using estimators such as Equation 7.43 will usually be inaccurate because of bias errors (e.g., observed temperatures warmer or cooler, on average, than the forecast temperatures), and/or dispersion errors (ensemble dispersion smaller or larger, on average, than required to accurately characterize the forecast uncertainty), which occur in general because of imperfect ensemble initialization and deficiencies in the structure of the dynamical model. Ordinary MOS postprocessing of single-integration dynamical forecasts through regression methods (Section 7.5.2) can be extended to compensate for ensemble dispersion errors also, by using an ensemble dispersion predictor in the regression. Adjusting the dispersion of the ensemble according to its historical error statistics can allow information on possible state-, or flow-dependent predictability to be included also in an ensemble MOS procedure.

One regression-based ensemble MOS approach that has been successful is logistic regression (Section 7.3.2) using the ensemble mean as one predictor, together with a second predictor involving the ensemble standard deviation. Wilks and Hamill (2007) used the formulation

$$\Pr\{V \le q\} = \frac{\exp(b_0 + b_1 \bar{x}_{ens} + b_2 \bar{x}_{ens} s_{ens})}{1 + \exp(b_0 + b_1 \bar{x}_{ens} + b_2 \bar{x}_{ens} s_{ens})}, \tag{7.44}$$

where $\bar{x}_{ens}$ is the ensemble mean and s_{ens} is the ensemble standard deviation. Another possible formulation, which yielded slightly better forecasts in an artificial data setting (Wilks, 2006b), is to specify the second predictor simply as the ensemble standard deviation rather than the product of the ensemble mean and ensemble standard deviation. However, Equation 7.44 has the appealing interpretation that it is equivalent to a logistic regression that uses the ensemble mean as the single predictor, but in which the regression parameter b_1 is itself a linear function of the ensemble standard deviation. Therefore, the steepness of the logistic function as it rises or falls with its characteristic S shape can increase with decreasing ensemble spread, yielding sharper forecasts (i.e., more frequent use of extreme probabilities) when the ensemble spread is small.

Figure 7.35 illustrates this idea for the case of 1-day ahead forecasts of January maximum temperature at Atlanta, Georgia. The predictand is the probability that the temperature will be at or below its 90th percentile, which is approximately 65°F. The forecast probability decreases as the ensemble mean maximum temperature forecast increases, and the decrease is steeper as the ensemble standard deviation decreases. The specific parameters for Equation 7.44 leading to the curves in Figure 7.35 are $b_0 = 15.2$, $b_1 = –0.245$, and $b_2 = 0.733$, which were fit on the basis of the performance of a particular set of ensemble forecasts that had been computed retrospectively for 25 years of historical weather using a fixed dynamical model (Hamill et al., 2006), called *reforecasts*. Figure 7.35 shows logistic curves for only three selected levels of ensemble standard deviation, but Equation 7.44 defines a continuum of these curves as a function of the ensemble standard deviation.

Experience to date has indicated that the second predictor in Equation 7.44, involving the ensemble standard deviation, may not be justified by the data in cases where the training sample size is small, or for relatively long lead times (Hamill et al., 2004; Wilks and Hamill, 2007). Since the parameters in

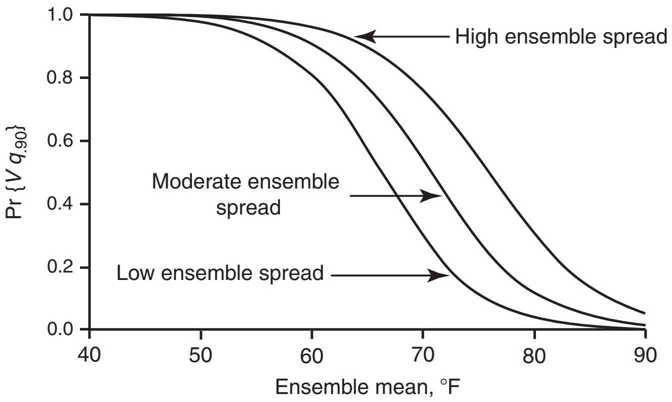


FIGURE 7.35 Logistic regressions in the form of Equation 7.44, for three selected levels of ensemble standard deviation. The predictand is probability of daily January Atlanta maximum temperature below its 90th percentile, which is approximately 65°F.

Equation 7.44 will generally be fit using maximum likelihood, whether the data justify $b_2 \neq 0$ can be evaluated using a likelihood ratio test (Section 5.2.6) or, if the possible use of additional predictors is to be evaluated also, the BIC (Equation 7.32) or the AIC (Equation 7.33) statistics.

One drawback to using logistic regressions such as Equation 7.44 for ensemble-MOS prediction of continuous quantities, whether or not an ensemble-spread predictor is included, is that separate equations are usually fit for each of a finite number of forecast quantiles. One consequence is that a large number of regression parameters must then be calculated, increasing the probability that some will be poorly estimated especially if the training sample size is limited. Another potential problem is that the different logistic regressions for different predictand quantiles may be mutually inconsistent, possibly leading to nonsense forecasts such as negative probabilities for some ranges of the predictand.

Figure 7.36b illustrates the latter problem, for probability forecasts of 5-day accumulated precipitation (lead time 6–10 days), for November 28 through December 2, at Minneapolis, Minnesota. Here seven separate logistic regressions of the form

$$\Pr\{V \leq q\} = \frac{\exp(b_0 + b_1\sqrt{\bar{x}_{ens}})}{1 + \exp(b_0 + b_1\sqrt{\bar{x}_{ens}})} \tag{7.45}$$

have been fit, one for each of the indicated quantiles, q, of the climatological distribution of 5-day accumulated precipitation for this location and time of year. The square root of the ensemble mean has been used as the predictor because it yields better forecasts for this positively skewed predictand. The ensemble spread has not been used because it did not significantly improve the predictions for this relatively long lead time. The main pathological feature of Figure 7.26b is that the regression lines cross on the log-odds scale, for $\bar{x}_{ens}$ largerthan about 3 mm (the point at which the regression functions for $q_{0.33}$ and $q_{0.50}$ intersect), implying that the resulting forecast probabilities overall would be incoherent. For example, when $\bar{x}_{ens} > 3$mm, the logistic regression for the median (2.03 mm) predicts smaller probabilities than does the logistic regression for the lower tercile (0.51 mm), which is clearly impossible.

This problem of potentially incoherent forecast probabilities can be avoided by fitting logistic regressions for all quantiles simultaneously, including an additional predictor that is a

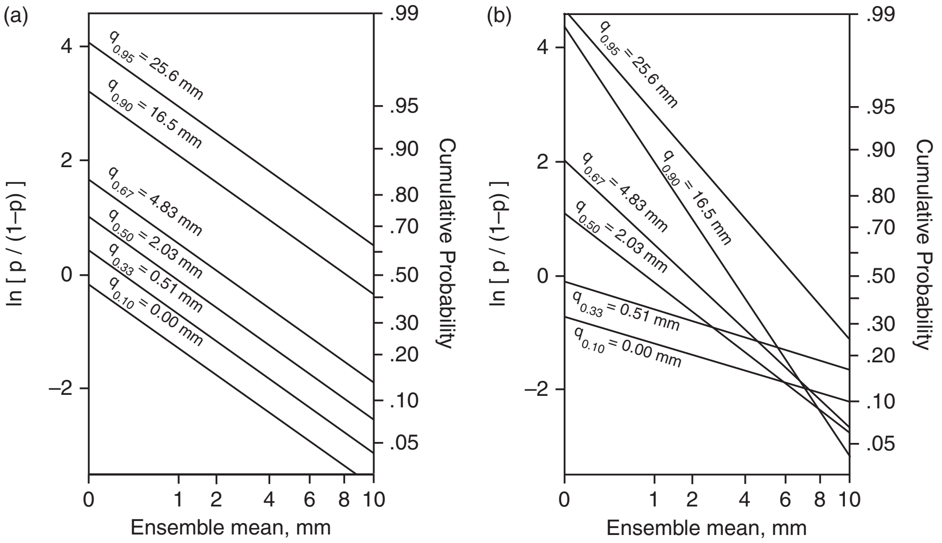


FIGURE 7.36 Logistic regressions plotted on the log-odds scale, for November 28 – December 2 accumulated precipitation at Minneapolis, at 6-10 day lead time. Forecasts from Equation (7.46), evaluated at selected quantiles, are shown by the parallel lines in panel (a), which cannot yield logically inconsistent sets of forecasts. Regressions for the same quantiles, fitted separately using Equation (7.45), are shown in panel (b). Because these regressions are not constrained to be parallel, logically inconsistent forecasts are inevitable for sufficiently extreme values of the predictor. *From Wilks (2009).*

(generally nonlinear) function, $g(q)$, of the forecast quantile itself, yielding the unified logistic regression model

$$\Pr\{V \leq q\} = \frac{\exp[g(q) + f(\bar{x}_{ens})]}{1 + \exp[g(q) + f(\bar{x}_{ens})]}. \tag{7.46}$$

When justified by the data, the predictor function $f(\bullet)$ would be extended to include a measure of ensemble spread also, for example $f(\bar{x}_{ens}, s_{ens}) = b_0 + b_1 \bar{x}_{ens} + b_2 s_{ens}$. Equation 7.45 is a special case of Equation 7.46, with $f(\bar{x}_{ens}) = b_0 + b_1 \sqrt{\bar{\mathrm{x}}_{\mathrm{ens}}}$, and $g(q) = 0$. The log-odds parallel regression lines in Figure 7.36a resulted from fitting the same data used in Figure 7.36b, simultaneously using Equation 7.46; with $f(\bar{x}_{ens}) = b_0 + b_1 \sqrt{\bar{\mathrm{x}}_{\mathrm{ens}}}$ as in Equation 7.45, but also $g(q) = b_2 \sqrt{q}$. The result is that all the forecast functions have log-odds slope b_1, and intercept $b_0 + b_2 \sqrt{q}$, for any forecast quantile q. Because these regression functions cannot cross, the resulting forecasts cannot yield incoherent probabilities. Additional advantages are that probabilities for any predictand quantile (or an entire forecast CDF) can be computed, and the number of parameters that must be estimated is greatly reduced. Fuller details are provided in Wilks (2009).

A different approach to computing MOS corrections to ensemble forecasts is based on an extension to linear regression (Section 7.2), but allowing the residual variance to depend linearly on the ensemble variance, yielding more uncertain (higher-variance) forecast distributions when the ensemble spread is large, and sharper (lower-variance) forecast distributions when ensemble spread is small. The method, proposed by Gneiting et al. (2005), is known as *nonhomogeneous Gaussian regression* (NGR) because the residual variance is allowed to be nonconstant (nonhomogeneous) from forecast to forecast, rather than being assumed equal for all predictions as in ordinary linear regression.

The usual, and simplest, formulation of NGR for ensemble-MOS applications consists of a simple linear regression using the ensemble mean as the only predictor,

$$V = a + b\bar{x}_{ens} + \varepsilon, \tag{7.47a}$$

where the variance of the residuals ε, which are assumed to have Gaussian distributions, is specified as a linear function of the ensemble variance,

$$\sigma_\varepsilon^2 = c + d\, s_{ens}^2. \tag{7.47b}$$

Equation 7.47a could also be extended to include more than the single predictor.

There are four parameters to be estimated in Equation 7.47—*a*, *b*, *c*, and *d*—but analytical solutions for them, analogous to those for simple linear regression in Section 7.2.1, are not available. Rather than estimating these parameters in a conventional way, for example, by maximizing their joint likelihood (Section 4.6) assuming Gaussian distributions for the residuals, Gneiting et al. (2005) also proposed the innovation of choosing the parameters to minimize the continuous ranked probability score (CRPS, Section 8.5.1), averaged over all forecasts in the training data set. Assuming that the forecast distribution will be Gaussian, the CRPS for a single postprocessed forecast characterized by the parameters $\mu = a + b\bar{x}_{ens}$ and $\sigma_\varepsilon{}^2 = c + ds^2{}_{ens}$, and its corresponding verifying observation (V) is

$$CRPS = \sigma_\varepsilon\left[z\,(2\Phi(z) - 1) + 2\phi(z) - \frac{1}{\sqrt{\pi}}\right], \tag{7.48a}$$

where

$$z = \frac{V - \mu}{\sigma_\varepsilon} \tag{7.48b}$$

is the observation standardized using its predicted value and the predicted residual variance, and $\phi(\bullet)$ and $\Phi(\bullet)$ denote, respectively, the PDF and CDF of the standard Gaussian distribution. Minimization of Equation 7.48 with respect to the four regression parameters requires use of iterative numerical methods.

Once the four parameters in Equation 7.47 have been estimated, probability forecasts are generated using

$$\Pr\{V \le q\} = \Phi\left[\frac{q - (a + b\bar{x}_{ens})}{(c + d\sigma_{ens}^2)^{1/2}}\right]. \tag{7.49}$$

Thus, in common with the unified logistic regression model in Equation 7.46, Equation 7.47 does not require a separate set of parameters to be estimated for each quantile q for which forecasts are needed. However, its use is appropriate only in situations where the distributions of regression residuals are reasonably represented by Gaussian PDFs.

Wilks and Hamill (2007), Hagedorn et al. (2008), and Kann et al. (2009) have reported good results when postprocessing ensemble forecasts for surface temperatures (which are approximately Gaussian) using NGR, yielding substantial improvements in forecast skill over direct use of raw ensemble output (e.g., Equation 7.42 or 7.43). Thorarinsdottir and Gneiting (2010) extend NGR to handle predictands such as wind speeds, forecasts for which must be non-negative.

Bremnes (2004) describes forecasts of probability distributions for precipitation using a two-stage ensemble MOS procedure based on *quantile regression*, with selected quantiles of the forecast ensemble precipitation distribution as predictors. First, the probability of nonzero precipitation is forecast using a *probit regression*, which is similar to logistic regression (Equation 7.29), but using the CDF of the standard Gaussian distribution to constrain the linear function of the predictors to the unit interval. That is, $p_i = \Phi(b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3)$, where the three predictors are the ensemble minimum, the ensemble median, and the ensemble maximum. Second, conditional on the occurrence of nonzero precipitation, the 5th, 25th, 50th, 75th, and 95th percentiles of the precipitation amount distributions are specified with separate regression equations, which each use the two ensemble quartiles as predictors. The final postprocessed precipitation probabilities then are obtained through the multiplicative law of probability (Equation 2.11), where E_1 is the event that nonzero precipitation occurs, and E_2 is a precipitation amount event defined by some combination of the forecast percentiles (e.g., the IQR) produced by the second regression step.

7.7.3. Kernel Density (Ensemble "Dressing") Methods

A quite different approach to estimating smooth forecast PDFs from a finite ensemble is provided through the use of kernel density methods (Section 3.3.6). In this approach, an individual PDF (kernel) is centered at each ensemble member, and the forecast PDF is then formed from a weighted average of this n_{ens}-member collection of kernels. In effect, each point ensemble forecast is "dressed" with a distribution that implies uncertainty around it, and the overall forecast PDF is then aggregated from the individual kernels. This is an ensemble MOS procedure because the distributions that are superimposed are derived from historical error statistics of the ensemble prediction system being postprocessed. Because individual ensemble members rather than the ensemble mean are dressed, the procedure yields state-dependent uncertainty information even if the spread of the added error distributions is not conditional on the ensemble spread.

Two critical issues must be addressed when constructing a kernel density estimate from a finite forecast ensemble. First, any bias in the historical performance of ensembles from the dynamical model being postprocessed must be removed at the outset. Often a simple, constant bias correction is applied to all ensemble members. For example, if the temperature forecasts from a particular dynamical model for a given location and season are too warm on average by 1°C (this would be estimated using a data set of historical forecasts and their corresponding observations), then on average each ensemble member will be too warm on average by 1°C, and this temperature bias would be subtracted from each ensemble member before computing the kernel density estimate. Bishop and Shanley (2008), Brocker and Smith (2008), Glahn et al. (2009b), and Unger et al. (2009) used linear regressions for de-biasing, so that, for example, ensemble systems for which low forecast temperatures are too warm but high forecast temperatures are too cool, can be accommodated. If forecast biases are not corrected before calculation of the kernel density estimate, the forecast PDF will be biased also. The second critical issue in ensemble dressing is choice of the mathematical form of the dressing kernel and, especially, its dispersion or bandwidth (the parameter h in Equation 3.13). The different ensemble dressing methods differ primarily with respect to how these two issues are addressed, although in all cases these characteristics are derived from historical ensemble forecast errors. Thus, ensemble dressing is indeed a MOS approach.

Most frequently, the dressing kernels are chosen to be Gaussian distributions, in which case the result is often called *Gaussian ensemble dressing* (GED). However, even though the dressing kernels are specified as Gaussian, the overall forecast distribution is in general not Gaussian; indeed, it can take on any shape that might be indicated by the distribution of the underlying ensemble members.

Ensemble dressing was first proposed by Roulston and Smith (2003), who derived dressing kernel characteristics (in particular, estimated the variance for the Gaussian dressing kernels) from the errors of the single best ensemble member (the member closest to the verifying observation) on a given forecast occasion. That is, the variance of each Gaussian dressing kernel is defined as the average squared difference between the observation and the (de-biased) ensemble member nearest to it, over a sample of past ensemble forecasts. This "best member" approach is conceptually appropriate because the dressing kernel should represent only uncertainty not already reflected in the ensemble dispersion.

Wang and Bishop (2005) proposed an alternative method for defining the Gaussian dressing variance, which they called second-moment constraint dressing. Working with de-biased ensembles so that the mean of each Gaussian dressing kernel is zero, according to this method the dressing variance is computed as

$$\sigma_D^2 = \sigma_{\bar{X}_{ens}-V}^2 - \left(1 + \frac{1}{n_{ens}}\right)\bar{\sigma}_{ens}^2. \tag{7.50}$$

Here the first term on the right is the sample variance for the errors of (the de-biased) ensemble-mean forecasts (i.e., their mean-squared error), and the second term is a slightly inflated average ensemble variance, again estimated from a history of like ensemble forecasts. Equation 7.50 provides an attractive alternative to best-member dressing when Gaussian dressing kernels are appropriate because it is not necessary to determine the best member of each ensemble in the training data, which process can be problematic in real forecast settings where the dimension of the dynamical model is quite large (Roulston and Smith, 2003). However, Equation 7.50 can sometimes fail (i.e., yield negative dressing variances) if the forecast ensembles in the training data are sufficiently overdispersed, on average. Even when Equation 7.50 yields a positive variance, if that variance is sufficiently small, the resulting forecast distribution may exhibit spurious multimodality, or "spikes" associated with each ensemble member (Bishop and Shanley, 2008). More generally, ensemble dressing methods are well suited to the usual condition of underdispersed ensembles because the dressing kernels add variance to the underlying dispersion of the ensemble members, but the method cannot reduce the variance of overdispersed ensembles.

Regardless of whether the Gaussian kernel variance ${\sigma_D}^2$ is estimated as the variance of best-member errors or using Equation 7.50, GED forecast probabilities are computed using

$$\Pr\{V \le q\} = \frac{1}{n_{ens}} \sum_{i=1}^{n_{ens}} \Phi\left[\frac{q - \tilde{x}_i}{\sigma_D}\right], \tag{7.51}$$

which weights all of the n_{ens} Gaussian kernels equally. The tilde over the ith ensemble member denotes that it has been de-biased, as indicated previously. Equations 7.50 and 7.51 are appropriate to scalar forecasts, but generalize easily to higher-dimensional forecasts as well (Wang and Bishop, 2005).

The method of *Bayesian model averaging* (BMA) (Raftery et al., 2005) is closely allied to best-member ensemble dressing. The differences are that the dressing kernels need not be the same for all ensemble members, and the estimation method for the kernel dispersion is different. When Gaussian kernels are used, each may have a different variance. In settings where a single dynamical model is used to integrate all ensemble members, it may be that the "control" integration (initialized from the best estimate of the initial condition) will have somewhat different statistical characteristics from the other ensemble members, which are mutually statistically indistinguishable. In that case. the dressing variance for the control member can be allowed to differ from that used with the ensemble members, and the control member may be weighted differently from the other ensemble members. These parameters (the two kernel variances and two weights) are estimated by maximizing the log-likelihood function

$$\ln(\Lambda) = -\sum_{i=1}^{n} \ln\left[w_1\, g(v_i | \tilde{x}_{1,i}, \sigma_1^2) + \sum_{j=2}^{n_{ens}} w_e\, g(v_i | \tilde{x}_{j,i}, \sigma_e^2) \right], \tag{7.52}$$

with respect to the weights w and the variances $\sigma^{\,2}$ over the n-forecast training data. Here $g(\bullet)$ indicate Gaussian-PDF kernels for the ith verification v_i centered on the de-biased ensemble members $\tilde{x}$, ${\sigma_1}^2$ is the dressing variance for the control member, and ${\sigma_e}^2$ is the dressing variance for the remaining ensemble members. The weights w_1 and w_e, with $w_1 + (n_{ens} - 1)w_e = 1$, allow unequal influence for the control member. Having estimated these parameters, BMA-based probability forecasts are computed, analogously to Equation 7.51, as

$$\Pr\{V \le q\} = w_1\, \Phi\left[\frac{q - \tilde{x}_1}{\sigma_1}\right] + \sum_{j=2}^{n_{ens}} w_e\, \Phi\left[\frac{q - \tilde{x}_j}{\sigma_e}\right]. \tag{7.53}$$

Bayesian model averaging is especially well suited to underdispersed *mult-model ensembles* (in which individual ensemble members or groups of ensemble members have been integrated using different dynamical models with different error characteristics), in which case Equations 7.52 and 7.53 can be extended to allow each group of ensemble members to have its own weight and dressing variance (e.g., Fraley et al., 2010).

Fortin et al. (2006) have proposed allowing different best-member dressing kernels for different ensemble members, depending on their rank within the ensemble. Bishop and Shanley (2008) and Fortin et al. (2006) note that ensemble dressing methods may overestimate probabilities for extremes events when the ensemble mean is far from the climatological average. Sloughter et al. (2007) describe BMA for precipitation forecasts, using a mixed discrete (representing the probability of zero precipitation) and continuous (gamma distribution representing nonzero precipitation amounts) kernel. Brocker and Smith (2008) extend ensemble dressing in a way that handles both overdispersed and underdispersed ensembles.

7.8. SUBJECTIVE PROBABILITY FORECASTS

7.8.1. The Nature of Subjective Forecasts

Most of this chapter has dealt with objective forecasts, or forecasts produced by means that are automatic. Objective forecasts are determined unambiguously by the nature of the forecasting procedure and the values of the variables that are used to drive it. However, objective forecasting procedures necessarily rest on a number of subjective judgments made during their development. Nevertheless, some people feel more secure with the results of objective forecasting procedures, seemingly taking comfort from their lack of contamination by the vagaries of human judgment. Apparently, such individuals feel that objective forecasts are in some way less uncertain than human-mediated forecasts.

One very important—and perhaps irreplaceable—role of human forecasters in the forecasting process is in the subjective integration and interpretation of objective forecast information. These objective forecast products often are called forecast guidance, and include deterministic forecast information from dynamical integrations, and statistical guidance from MOS systems or other interpretive statistical products. Human forecasters also use, and incorporate into their judgments, available atmospheric observations (surface maps, radar images, etc.), and prior information ranging from persistence or simple climatological statistics, to their individual previous experiences with similar meteorological situations. The result is (or should be) a forecast reflecting, to the maximum practical extent, the forecaster's state of knowledge about the future evolution of the atmosphere.

Human forecasters can rarely, if ever, fully describe or quantify their personal forecasting processes (Stuart et al., 2007). Thus, the distillation by a human forecaster of disparate and sometimes conflicting information is known as *subjective* forecasting. A subjective forecast is one formulated on the basis of the judgment of one or more individuals. Making a subjective weather forecast is a challenging process precisely because future states of the atmosphere are inherently uncertain. The uncertainty will be larger or smaller in different circumstances—some forecasting situations are more difficult than others—but it will never really be absent. Doswell (2004) provides some informed perspectives on the formation of subjective judgments in weather forecasting.

Since the future states of the atmosphere are inherently uncertain, a key element of a good and complete subjective weather forecast is the reporting of some measure of the forecaster's uncertainty. It is the forecaster who is most familiar with the atmospheric situation, and it is therefore the forecaster who is in the best position to evaluate the uncertainty associated with a given forecasting situation. Although it is common for nonprobabilistic forecasts (i.e., forecasts containing no expression of uncertainty) to be issued, such as "tomorrow's maximum temperature will be 27°F," an individual issuing this forecast would not seriously expect the temperature to be exactly 27°F. Given a forecast of 27°F, temperatures of 26 or 28°F would generally be regarded as nearly as likely, and in this situation the forecaster would usually not really be surprised to see tomorrow's maximum temperature anywhere between 25 and 30°F.

Although uncertainty about future weather can be reported verbally using phrases such as "chance" or "likely," such qualitative descriptions are open to different interpretations by different people (e.g., Murphy and Brown, 1983). Even worse, however, is the fact that such qualitative descriptions do not precisely reflect the forecasters' uncertainty about, or degree of belief in, the future weather. The forecaster's state of knowledge is most accurately reported, and the needs of the forecast user are best served, if the intrinsic uncertainty is quantified in probability terms. Thus, the Bayesian view of probability as the degree of belief of an individual holds a central place in subjective forecasting. Note that since different forecasters have somewhat different information on which to base their judgments (e.g., different sets of experiences with similar past forecasting situations), it is perfectly reasonable to expect that their probability judgments may differ somewhat as well.

7.8.2. The Subjective Distribution

Before a forecaster reports a subjective degree of uncertainty as part of a forecast, he or she needs to have a mental image of that uncertainty. The information about an individual's uncertainty can be thought of as residing in the individual's *subjective distribution* for the event in question. The subjective distribution is a probability distribution in the same sense as the parametric distributions described in Chapter 4. Sometimes, in fact, one of the distributions specifically described in Chapter 4 may provide a very good approximation to an individual's subjective distribution. Subjective distributions are interpreted from a Bayesian perspective as the quantification of an individual's degree of belief in each of the possible outcomes for the variable being forecast.

Each time a forecaster prepares to make a forecast, he or she internally develops a subjective distribution. The possible weather outcomes are subjectively weighed, and an internal judgment is formed as to their relative likelihoods. This process occurs whether or not the forecast is to be a probability forecast, or indeed whether or not the forecaster is even consciously aware of the process. However, unless we believe that uncertainty can somehow be expunged from the process of weather forecasting, it should be clear that better forecasts will result when forecasters think explicitly about their subjective distributions and the uncertainty that those distributions describe.

It is easiest to approach the concept of subjective probabilities with a familiar but simple example. Subjective probability-of-precipitation (PoP) forecasts have been routinely issued in the United States since 1965. These forecasts specify the probability that measurable precipitation (i.e., at least 0.01 in.) will occur at a particular location during a specified time period. The forecaster's subjective distribution for this event is so simple that we might not notice that it is a probability distribution. However, the events "precipitation" and "no precipitation" divide the sample space into two MECE events. The distribution of probability over these events is discrete and consists of two elements: one probability for the event "precipitation" and another probability for the event "no precipitation." This distribution will be different for different forecasting situations, and perhaps for different forecasters assessing the same situation. However, the only thing about a forecaster's subjective distribution for the PoP that can change from one forecasting occasion to another is the probability, and this will be different to the extent that the forecaster's degree of belief regarding future precipitation occurrence is different. The PoP ultimately issued by the forecaster should be the forecaster's subjective probability for the event "precipitation," or perhaps a suitably rounded version of that probability. That is, it is the forecaster's job to evaluate the uncertainty associated with the possibility of future precipitation occurrence and to report that uncertainty to the users of the forecasts.

7.8.3. Central Credible Interval Forecasts

It has been argued here that inclusion of some measure of the forecaster's uncertainty should be included in any weather forecast. Forecast users can use the added uncertainty information to make better, economically more favorable, decisions (e.g., Roulston et al., 2006). Historically, resistance to the idea of probability forecasting has been based in part on the practical consideration that the forecast format should be compact and easily understandable. In the case of PoP forecasts, the subjective distribution is sufficiently simple that it can be reported with a single number, and it is no more cumbersome than issuing a nonprobabilistic forecast of "precipitation" or "no precipitation." When the subjective distribution is continuous, however, some approach to sketching its main features is a practical necessity if its probability information is to be conveyed succinctly in a publicly issued forecast. Discretizing a continuous subjective distribution is one approach to simplifying it in terms of one or a few easily expressible quantities. Alternatively, if the forecaster's subjective distribution on a given occasion can be reasonably well approximated by one of the parametric distributions described in Chapter 4, another approach to simplifying its communication could be to report the parameters of the approximating distribution. There is no guarantee, however, that subjective distributions will always (or even ever) correspond to a familiar parametric form.

One very attractive and workable alternative for introducing probability information into forecasts for continuous meteorological variables is the use of *credible interval forecasts*. This forecast format has been used operationally in Sweden (Ivarsson et al., 1986), but to date has been used only experimentally in the United States (Murphy and Winkler, 1974; Peterson et al., 1972; Winkler and Murphy, 1979). In unrestricted form, a credible interval forecast requires specification of three quantities: two points defining an interval for the continuous forecast variable, and a probability (according to the forecaster's subjective distribution) that the forecast quantity will fall in the designated interval. Usually the requirement is also made that the credible interval be located in the middle of the subjective distribution. In this case the specified probability is distributed equally on either side of the subjective median, and the forecast is called a *central credible interval* forecast.

There are two special cases of the central credible interval forecast format, each requiring that only two quantities be communicated. The first is the fixed-width central credible interval forecast. As the name implies, the width of the central credible interval is the same for all forecasting situations and is specified in advance for each predictand. Thus the forecast includes a location for the interval, generally specified as its midpoint, and a probability that the outcome will occur in the forecast interval. For example, the Swedish central credible interval forecasts for temperature are of the fixed-width type, with the interval size specified to be ±3°C around the midpoint temperature. These forecasts thus include a forecast temperature, together with a probability that the subsequently observed temperature will be within 3°C of the forecast temperature. The two forecasts 15°C, 90% and 15°C, 60% would both indicate that the forecaster expects the temperature to be about 15°C, but the inclusion of probabilities in the forecasts shows that much more confidence can be placed in the former as opposed to the latter of the two forecasts of 15°C. Because the forecast interval is central, these two forecasts would also imply 5% and 20% chances, respectively, for the temperature to be colder than 12° or warmer than 18°.

Some forecast users would find the unfamiliar juxtaposition of a temperature and a probability in a fixed-width central credible interval forecast to be somewhat jarring. An alternative forecast format that could be implemented more subtly is the fixed-probability central credible interval forecast. In this format, it is the probability contained in the forecast interval, rather than the width of the interval, that is specified in advance and is constant from forecast to forecast. This format makes the probability component of the credible interval forecast implicit, so the forecast consists of two numbers having the same physical dimensions as the quantity being forecast.

Figure 7.37 illustrates the relationship of 75% central credible intervals for two subjective distributions having the same mean. The shorter, broader distribution represents a relatively uncertain forecasting situation, where events fairly far away from the center of the distribution are regarded as having substantial probability. A relatively wide interval is therefore required to subsume 75% of this distribution's probability. On the other hand, the tall and narrow distribution describes considerably less uncertainty, and a much narrower forecast interval contains 75% of its density. If the variable being forecast is temperature, the 75% central credible interval forecasts for these two cases might be 10° to 20° and 14° to 16°, respectively.

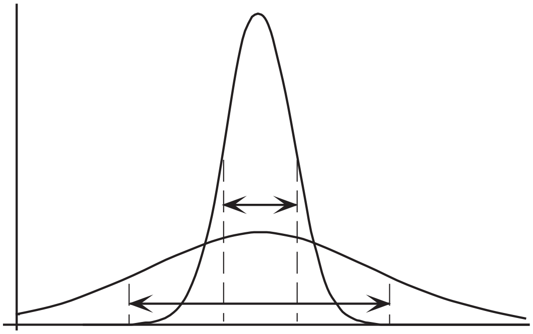


FIGURE 7.37 Two hypothetical subjective distributions shown as probability density functions. The two distributions have the same mean, but reflect different degrees of uncertainty. The tall, narrow distribution represents an easier (less uncertain) forecasting situation, and the broader distribution represents a more difficult forecast problem. Arrows delineate 75% central credible intervals in each case.

A strong case can be made for operational credible interval forecasts (Murphy and Winkler, 1974, 1979). Since nonprobabilistic temperature forecasts are already often specified as ranges, fixed-probability central credible interval forecasts could be introduced into forecasting operations quite unobtrusively. Forecast users not wishing to take advantage of the implicit probability information would notice little difference from the present forecast format, whereas those understanding the meaning of the forecast ranges would derive additional benefit. Even forecast users unaware that the forecast range is meant to define a particular interval of fixed probability might notice over time that the interval widths were related to the precision of the forecasts.

7.8.4. Assessing Discrete Probabilities

Experienced weather forecasters are able to formulate subjective probability forecasts that evidently quantify their uncertainty regarding future weather quite successfully. Examination of the error characteristics of such forecasts (see Chapter 8) reveals that they are largely free of the biases and inconsistencies sometimes exhibited in the subjective probability assessments made by less experienced individuals. Commonly, inexperienced forecasters produce probability forecasts exhibiting overconfidence (Murphy, 1985), or biases due to such factors as excessive reliance on recently acquired information (Spetzler and Staël von Holstein, 1975; Tversky, 1974).

Individuals who are experienced in assessing their subjective probabilities can do so in a seemingly unconscious or automatic manner. People who are new to the practice often find it helpful to use physical or conceptual devices that allow comparison of the uncertainty to be assessed with an uncertain situation that is more concrete and familiar (Garthwaite et al., 2005). For example, Spetzler and Staël von Holstein (1975) describe a physical device called a probability wheel, which consists of a spinner of the sort that might be found in a child's board game, on a background that has the form of a pie chart. This background has two colors, blue and orange, and the proportion of the background covered by each of the colors can be adjusted. The probability wheel is used to assess the probability of a dichotomous event (e.g., a PoP forecast) by adjusting the relative coverages of the two colors until the forecaster feels the probability of the event to be forecast is about equal to the probability of the spinner stopping in the orange sector. The subjective probability forecast is then read as the angle subtended by the orange sector, divided by 360°.

Conceptual devices can also be employed to assess subjective probabilities. For many people, comparison of the uncertainty surrounding the future weather is most easily assessed in the context of lottery games or betting games. Such conceptual devices translate the probability of an event to be forecast into more concrete terms by posing hypothetical questions such as "would you prefer to be given $2 if precipitation occurs tomorrow, or $1 for sure (regardless of whether or not precipitation occurs)?" Individuals preferring the sure $1 in this lottery situation evidently feel that the relevant PoP is less than 0.5, whereas individuals who feel the PoP is greater than 0.5 would generally prefer to receive $2 on the chance of precipitation. A forecaster can use this lottery device by adjusting the variable payoff relative to the certainty equivalent (the sum to be received for sure) until the point of indifference, where either choice would be equally attractive. That is, the variable payoff is adjusted until the expected (i.e., probability-weighted average) payment is equal to the certainty equivalent. Denoting the subjective probability as p, the procedure can be written formally as

$$\text{Expected payoff} = p\,(\text{Variable payoff}) + (1-p)(\ \$0) = \text{Certainty equivalent} \tag{7.54a}$$

which leads to

$$p = \frac{\text{Certainty equivalent}}{\text{Variable payoff}}. \tag{7.54b}$$

The same kind of logic can be applied in an imagined betting situation. Here the forecasters ask themselves whether receiving a specified payment should the weather event to be forecast occurs, or suffering some other monetary loss if the event does not occur, is preferable. In this case the subjective probability is assessed by finding monetary amounts for the payment and loss such that the bet is a fair one, implying that the forecaster would be equally happy to be on either side of it. Since the expected payoff from a fair bet is zero, the betting game situation can be represented as

$$\text{Expected payoff} = p\ (\ \$\ \text{payoff}) + (1-p)(-\ \ \$\ \text{loss}) = 0, \tag{7.55a}$$

leading to

$$p = \frac{\$\ \text{loss}}{\$\ \text{loss} +\ \$\ \text{payoff}}. \tag{7.55b}$$

Many betting people think in terms of odds in this context. Equation 7.55a can be expressed alternatively as

$$\text{odds ratio} = \frac{p}{1-p} = \frac{\$\ \text{loss}}{\$\ \text{payoff}}. \tag{7.56}$$

Thus, a forecaster being indifferent to an even-money bet (1:1 odds) harbors an internal subjective probability of $p = 0.5$. Indifference to being on either side of a 2:1 bet implies a subjective probability of 2/3, and indifference at 1:2 odds is consistent with an internal probability of 1/3.

7.8.5. Assessing Continuous Distributions

The same kinds of lotteries or betting games just described can also be used to assess quantiles of a subjective continuous probability distribution using the *method of successive subdivision*. Here the approach is to identify quantiles of the subjective distribution by comparing event probabilities that they imply with the reference probabilities derived from conceptual money games. Use of this method in an operational setting is described in Krzysztofowicz et al. (1993).

The easiest quantile to identify is the median. Suppose the distribution to be identified is for tomorrow's maximum temperature. Since the median divides the subjective distribution into two equally probable halves, its location can be assessed by evaluating a preference between, say, \$1 for sure and \$2 if tomorrow's maximum temperature is warmer than 14°C. The situation is the same as that described in Equation 7.54. Preferring the certainty of \$1 implies a subjective probability for the event {maximum temperature warmer than 14°C} that is smaller than 0.5. A forecaster preferring the chance at \$2 evidently feels that the probability for this event is larger than 0.5. Since the cumulative probability, p, for the median is fixed at 0.5, we can locate the threshold defining the event {outcome above median} by adjusting it to the point of indifference between the certainty equivalent and a variable payoff equal to twice the certainty equivalent.

The quartiles can be assessed in the same way, except that the ratios of certainty equivalent to variable payoff must correspond to the cumulative probabilities of the quartiles; that is, 1/4 or 3/4. At what temperature T_{LQ} are we indifferent to the alternatives of receiving \$1 for sure, or \$4 if tomorrow's maximum temperature is below T_{LQ}? The temperature T_{LQ} then estimates the forecaster's subjective lower quartile. Similarly, the temperature T_{UQ}, at which we are indifferent to the alternatives of \$1 for sure or \$4 if the temperature is above T_{UQ}, estimates the upper quartile.

Especially when someone is inexperienced at probability assessments, it is a good idea to perform some consistency checks. In the method just described, the quartiles were assessed independently, but together define a range—the 50% central credible interval—in which half the probability should lie. Therefore a good check on their consistency would be to verify that we are indifferent to the choices between \$1 for sure and \$2 if $T_{\mathrm{LQ}} \leq T \leq T_{\mathrm{UQ}}$. If we prefer the certainty equivalent in this comparison, the quartile estimates T_{LQ} and T_{UQ} are apparently too close. If we prefer the chance at the \$2, they apparently subtend too much probability. Similarly, we could verify indifference between the certainty equivalent, and four times the certainty equivalent if the temperature falls between the median and one of the quartiles. Any inconsistencies discovered in checks of this type indicate that some or all of the previously estimated quantiles need to be reassessed.

7.9. EXERCISES

7.1. a. Derive a simple linear regression equation using the data in Table A.3, relating June temperature (as the predictand) to June pressure (as the predictor).
b. Explain the physical meanings of the two parameters.
c. Formally test whether the fitted slope is significantly different from zero.
d. Compute the R^2 statistic.
e. Estimate the probability that a predicted value corresponding to $x_0 = 1013$ mb will be within 1°C of the regression line, using Equation 6.22.
f. Repeat (e), assuming the prediction variance equals the MSE.

7.2. Consider the following ANOVA table, describing the results of a regression analysis:

Source	df	SS	MS	F
Total	23	2711.60		
Regression	3	2641.59	880.53	251.55
Residual	20	70.01	3.50	

a. How many predictor variables are in the equation?
b. What is the sample variance of the predictand?
c. What is the R^2 value?
d. Estimate the probability that a prediction made by this regression will be within ± 2 units of the actual value.

7.3. Derive an expression for the maximum-likelihood estimate of the intercept b_0 in logistic regression (Equation 7.29), for the constant model in which $b_1 = b_2 = \ldots = b_K = 0$.

7.4. The 19 nonmissing precipitation values in Table A.3 can be used to fit the regression equation:

$$\ln[(\text{Precipitation}) + 1\ \text{mm}] = 499.4 - 0.512(\text{Pressure}) + 0.796(\text{Temperature})$$

The MSE for this regression is 0.701. (The constant 1 mm has been added to ensure that the logarithm is defined for all data values.)

a. Estimate the missing precipitation value for 1956 using this equation.
b. Construct a 95% prediction interval for the estimated 1956 precipitation.

7.5. Explain how to use cross validation to estimate the prediction mean squared error, and the sampling distribution of the regression slope, for the problem in Exercise 7.1. If the appropriate computing resources are available, implement your algorithm.

7.6. Hurricane Zeke is an extremely late storm in a very busy hurricane season. It has recently formed in the Caribbean, the 500-mb height at gridpoint 37 (relative to the storm) is 5400 m, the 500-mb height at gridpoint 3 is 5500 m, and the 1000-mb height at gridpoint 51 is –200 m (i.e., the surface pressure near the storm is well below 1000 mb).

a. Use the NHC 67 model (see Table 7.7) to forecast the westward component of its movement over the next 12 hours, if the storm has moved 80 n.mi. due westward in the previous 12 hours.

b. What would the NHC 67 forecast of the westward displacement be if, in the previous 12 hours the storm had moved 80 n.mi. westward *and* 30 n.mi. northward (i.e., $P_y = 30$ n. mi.)?

7.7. The fall (September, October, November) MOS equation for predicting maximum temperature (in °F) at Binghamton, New York, formerly used with a now-discontinued dynamical model, at the 60-hour lead time was

MAX T = –363.2 + 1.541 (850 mb T) – .1332 (SFC-490 mb RH) – 10.3 (COS DOY)

where:

(850 mb T) is the 48-hour dynamical forecast of temperature (K) at 850 mb
(SFC-490 mb RH) is the 48-hour forecast lower tropospheric RH in %
(COS DOY) is the cosine of the day of the year transformed to radians or degrees; that is,
= cos (2πt/365) or = cos (360° t / 365)
and *t* is the day number of the valid time (the day number for January 1 is 1, and for October 31 it is 304)

Calculate what the 60-hour MOS maximum temperature forecast would be for the following:

	Valid time	48-hr 850 mb T fcst	48-hr mean RH fcst
a.	September 4	278 K	30%
b.	November 28	278 K	30%
c.	November 28	258 K	30%
d.	November 28	278 K	90%

7.8. A MOS equation for 12–24 hour PoP in the warm season might look something like:

PoP = 0.25 + .0063(Mean RH) – .163(0-12 ppt [bin @ 0.1 in.]) – .165(Mean RH [bin @ 70%])

where:

Mean RH (%) is the same variable as in Exercise 7.7 for the appropriate lead time
0-12 ppt is the model-forecast precipitation amount in the first 12 hours of the forecast
[bin @ xxx] indicates use as a binary variable: =1 if the predictor is ≤ xxx
= 0 otherwise

Evaluate the MOS PoP forecasts for the following conditions:

	12-hour mean RH	0-12 ppt
a.	90%	0.00 in.
b.	65%	0.15 in.
c.	75%	0.15 in.
d.	75%	0.09 in.

7.9. Explain why the slopes of the solid lines decrease, from Figure 7.20 to Figure 7.21a, to Figure 7.21b. What would the corresponding MOS equation be for an arbitrarily long lead time into the future?

7.10. A forecaster is equally happy with the prospect of receiving $1 for sure, or $5 if freezing temperatures occur on the following night. What is the forecaster's subjective probability for frost?

7.11. A forecaster is indifferent between receiving $1 for sure and any of the following: $8 if tomorrow's rainfall is greater than 55 mm, $4 if tomorrow's rainfall is greater than 32 mm, $2 if tomorrow's rainfall is greater than 12 mm, $1.33 if tomorrow's rainfall is greater than 5 mm, and $1.14 if tomorrow's precipitation is greater than 1 mm.

a. What is the median of this individual's subjective distribution?

b. What would be a consistent 50% central credible interval forecast? A 75% central credible interval forecast?

c. In this forecaster's view, what is the probability of receiving more than one but no more than 32 mm of precipitation?